PNAS
www.pnas.org

Supplementary Information for
The genetic structure of the Turkish population reveals high levels of variation and admixture

M. Ece Kars, A. Nazlı Başak, O. Emre Onat, Kaya Bilguvar, Jungmin Choi, Yuval Itan, Caner Çağlar, Robin Palvadeau, Jean-Laurent Casanova, David N. Cooper, Peter D. Stenson, Alper Yavuz, Hakan Buluş, Murat Günel, Jeffrey M. Friedman, Tayfun Özçelik*

* Tayfun Özçelik
Email: tozcelik@bilkent.edu.tr

This PDF file includes:

Supplementary text
Figures S1 to S19
Tables S1 to S7
Legends for Datasets S1 to S6
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S6

Supplementary Methods and Materials

1. Study samples

Study samples comprised 3,864 TR individuals who either yielded whole exome (WES, *n* = 3,072) or whole genome (WGS, *n* = 792) sequence data, which were collected through different projects related to the molecular bases of human genetic disease (Table S1). We excluded 206 variants in the genes that were causally associated with the phenotypes in our cohort (Dataset S6). Written informed consent was obtained from all study participants during the sampling process for each study. All informed consents provided the permission to use the DNA samples and basic demographic information for disease gene identification studies and to share the data.

2. Sequencing and filtering

WES was performed at the Yale Center for Genome Analysis, TUBITAK or Macrogen using IDT xGen Exome Research 392 Panel v1.0 capture, Roche SeCap EZ Whole Exome V1, xGen Exome Research 392 Panel v1.0 capture, Roche SeCap EZ Whole Exome V2, xGen Exome Research 392 Panel v1.0 capture, Roche SeCap EZ Whole Exome V3 or Agilent SureSelect Human All Exon V6 kits according to the manufacturer's protocol. Samples were sequenced on the HiSeq2000, HiSeq2500 or HiSeq4000 platforms with 100-bp paired end-reads. The Illumina processing pipeline was used for base calling, read filtering, and demultiplexing. The read pairs were mapped to the human genome build GRCh37 using Burrows-Wheeler Aligner (BWA) v.0.7.17 (1). Duplicate reads were marked using Mark Duplicates tool in Picard tools. Base quality score recalibration (BQSR) and local realignment around indels were carried out with Genome Analysis Toolkit v.3.7 (GATK) (2). Variant discovery was performed following Best Practices workflows of GATK. HaplotypeCaller was employed to call variants, followed by joint genotyping using GenotypeGVCFs and splitting multiallelic variants with LeftAlignAndTrimVariants. To remove batch effects from the WES data, genotype calling was limited to the intersection of target regions of exome sequencing kits that overlap with consensus coding sequence (CCDS) build 15 coding exons (3).

WGS was performed on the Illumina HiSeq 2500 platform using PCR-free library preparation and 100-bp paired-end sequencing. Reads were aligned to the hg19 human genome build using BWA. The variants were called by the Isaac variant caller. The gVCF files for all WGS samples were jointly genotyped using Illumina gvcfgenotyper. Normalization, realignment around indels and splitting multiallelic variants were performed using BCFtools. The final joint VCF file was lifted over to human genome build GRCh37 with Picard tools using hg19 to b37 chain file, which was downloaded from UCSC website. We identified 2,694,125 WES and 72,982,375 WGS variants.

Statistical outliers of WES and WGS samples were evaluated separately using BCFtools stats. After the filtration according to the number of singletons, transition/transversion ratio, average depth and total number of variants, 89 WES samples were removed from the dataset because they fell outside five absolute deviations from the median. We did not identify any low-quality samples for WGS batch (Fig. S1 and S2).

For the selection of high-quality variants from the WES and WGS data, we used the following thresholds: a) Variants with Phred-scaled quality score < 30, b) genotypes with depth (DP) < 8, c) genotype quality < 20, and d) a missingness rate higher than 20% across all samples. In addition, variant quality score recalibration (VQSR) was performed for WES samples as implemented in GATK VariantRecalibrator. Variant recalibration was applied by ApplyRecalibration walker of GATK using tranche sensitivity of 99.5% for SNPs and 99.0% for indels. VQSR was used to define low quality variants for downstream processing. We restricted the analyses to genomic regions called by gnomAD and excluded all sites failing QC according to filtering method of gnomAD (*n* = 1,244,833) (4). We calculated the sample-based quality control measures and the mean number of novel variants (not present in dbSNP build 151) for the exome regions of the WES and WGS datasets using VariantEval walker of GATK (Table S3) (5).

We produced 38 technical replicates of which 11 were whole-exome sequenced in two different batches, 6 were whole-genome sequenced in two different batches and 21 were both whole-exome and whole-genome sequenced. We calculated the concordance rates of these

replicates using VariantEval tool of GATK after applying our QC filtering method (Table S7). A similar analysis was performed by GenomeAsia 100K study (6). We detected, on average, highly accurate genotypes with similar sensitivity, positive predictive value, and concordance rates to the previous studies investigating technical accuracy of sequencing data (7, 8).

Relatedness analysis was performed using KING and a kinship coefficient threshold 0.0884 was used to exclude second degree or closer relatives (9). Since the high level of inbreeding in the TR population is expected to result in an over-estimation of relatedness, we took a liberal approach and excluded the first and second-degree relatives (5, 9, 10). 413 samples were removed after this step. Finally, 773 WGS and 2,589 WES samples corresponding to a total of 3,362 individuals and 1,123,248 WES and 45,981,720 WGS variants constituted the downstream population structure and variome characterization studies.

Coverage calculations of the WES and WGS samples were performed using mosdepth, SAMtools and BCFtools. The mean target base coverage for the exons of CCDS build 15 of the WES samples was 70X with 95.32%, 93.81%, and 88.45% coverage at 8X, 10X and 20X or more, respectively. The mean depth of coverage for the WGS samples was 34X with 93.9%, 93.7% and 93.4% coverage at 8X, 10X and 20X or more, respectively.

The GRCh38 positions of the variants were obtained with Picard tools using the hg19 to GRCh38 chain file, which was downloaded from the UCSC website. 46,420,067 (99.3%) variants were successfully lifted over to GRCh38.

3. Population structure analyses

We produced four different datasets to evaluate the genetic structure of the TR population in a regional and global context. Populations used in the analyses were listed in Table S4. Exome data provides accurate results for population structure analyses (11). Therefore, the intersection of the target regions of the kits that were used during exome sequencing and CCDS regions were selected from TR WGS data and combined with TR WES data (n = 3,362) for the analysis of genetic variation within the TR population (TR dataset). Second, we selected 13 populations from 1000GP (12): African populations YRI and LWK; European populations GBR, TSI, IBS, and FIN; South Asian populations GIH, BEB, PJL, and ITU; East Asian populations CHB, CHS, and JPT (n = 1,299). We generated the "global dataset" by combining the data of the 1000GP populations and TR-WGS samples (n = 773) with that of Near-East populations from Lazaridis et al. (13) (n = 1,430). Third, we produced the "regional dataset" (n = 1,805) by listing the populations with the closest relationship with the TR population according to Wright's fixation index (F_{ST}). Lastly, we generated the "phylogeny dataset" (n = 5,357) by combining the allele frequency data of all TR samples, Middle Eastern populations from Scott et al. (14), and 1000GP to conduct phylogenetic tree analysis.

For all four datasets, SNVs were extracted from the VCF files using BCFtools and converted to PLINK binary file format. The sequencing data of Lazaridis et al. (13) was also converted to PLINK binary file format using the convertf utility of EIGENSOFT(15). The binary files were merged and variants were filtered using PLINK v.1.9 according to missingness (> 20%), deviation from Hardy-Weinberg equilibrium with a p-value of < 0.00005, minor allele frequency (MAF < 0.05), and linkage disequilibrium (r2 = 0.5) (16). All plots were generated with the aid of ggplot2, reshape, dplyr and stringr packages of R software (17-20).

***3.1. Origin of alleles*:** Grand-maternal and grand-paternal birthplace of the 1,460 (43.4%) individuals were obtained from patient records and the numbers of chromosomes from each region were depicted on a map of Turkey.

***3.2. Principal components analysis (PCA)*:** EIGENSOFT SmartPCA tool was used to demonstrate the degree of genetic variation between the populations (15). Three different PCAs were performed: The first was to explore the variation in Turkey using the origin-known TR samples from the TR dataset. The second was to explore the variation in a global context by using the global dataset, and the third was to display the relationship of the TR individuals with other populations in a regional context by using the regional dataset.

3.3. Procrustes analysis: Two symmetric Procrustes analyses with 100,000 permutations

were performed to evaluate the relationship of geographical distribution and genetic similarity of the TR and other populations. The values of the first two PCs of the PCA estimated using the origin-known TR samples from the TR dataset and the values of the first two PCs of the PCA estimated using the regional dataset were employed in the Procrustes analyses. The unprojected geographic coordinates (latitude-longitude) of the TR subregions were determined using geographical midpoints on the map of Turkey. The latitude and longitude of the capital cities were used for the other populations.

***3.4. Admixture*:** Substructures of the populations were assessed using ADMIXTURE (21). Analysis with *k* from 2 to 8 was run for the TR dataset in which *k* = 4 resulted in the lowest cross-validation error. Analysis with *k* from 2 to 14 was also run for the global dataset. The cross-validation error was lowest when 12 ancestral populations were present.

3.5. Phylogenetic tree: Population splitting and genetic drift in the populations included in the phylogeny dataset were evaluated by a maximum likelihood phylogenetic tree using Treemix software (22).

3.6. Wright's fixation index: The degree of differentiation among the populations included in the regional dataset was evaluated with F_{ST} values produced by Weir and Cockerham estimation, which is included in the EIGENSOFT SmartPCA.

3.7. Linkage disequilibrium (LD) decay: LD decay for the populations in the global dataset was calculated using PLINK. PLINK --r2 option with 70 kb sliding window and no limit for r2 was used to calculate pairwise correlations; they were binned by genomic distance between the SNPs (up to 70 kb), and averages were calculated for each bin.

3.8. Inbreeding coefficient: PLINK --het algorithm was used to determine the inbreeding coefficients (F_{plink}) of the individuals in the global dataset. We detected several individuals with negative F_{plink} values, which could reflect a recent admixture of previously diverse populations or biased variant sampling (23). We also listed the reported parental relationships of the TR WGS samples (538 unrelated, 56 endogamous, 95 consanguineous, and 84 unknown) and evaluated their effect on the inbreeding coefficient and runs of homozygosity (ROH) analyses.

3.9. Runs of homozygosity: Autosomal SNPs of unrelated TR WGS samples and WGS data of 1000GP samples were used to detect runs of homozygosity. SNPs with minor allele frequencies lower than 0.05 and those that diverted from Hardy-Weinberg equilibrium with $p < 0.00005$ were removed (24). The lengths of homozygous regions were calculated using PLINK --homozyg option. With a 50 SNP-containing 50 kb sliding window, ROH longer than 300 kb in length were determined. Three heterozygous calls were allowed during the analysis (24). We classified ROHs according to their length by using previously published ranges, as short (< 0.515 Mb), medium length (0.516–1.606 Mb), and long (> 1.607 Mb) (14, 25). We determined the proportion of the autosomal genome in runs of homozygosity above a specified length threshold (F_{roh}) (26):

$$F_{roh} = \sum L_{roh} / L_{auto}$$

where $\sum L_{roh}$ is the sum of an individual's ROHs above a specified threshold and L_{auto} is the total length of the autosomal genome except centromeres. We determined L_{auto} as 2,643,316 kb based on the percentage of coverage of the TR-WGS samples at 8X.

3.10. Y-chromosome and mitochondrial DNA (mtDNA) haplogroups: Y-chromosome haplotypes were inferred from the Global dataset using Y-Lineage Tracker based on the International Society of Genetic Genealogy (ISOGG) Y-DNA tree (2019 version) (27). The analysis was performed using the variants within the 10.3 Mb of the Y-chromosome (12). To assign TR samples to known mtDNA haplogroups, we aligned mtDNA variants to Revised Cambridge Reference Sequence (rCRS) of the human mtDNA (NC_012920). To assess the

mtDNA haplogroup diversity of the TR population in a global context, we downloaded mtDNA variants of 1000GP and Human Genome Diversity Project samples (28). Haplogroups were assigned using Haplogrep2 (v2.2) and Phylotree v17 as a reference (29, 30). Only the major Y-chromosome and mtDNA haplogroups were shown in the Fig. S12-S15. High-resolution assessments were listed in the Dataset S1. Central Asian specific Y-chromosome haplotypes were previously identified as C-RPS4Y and O3-M122. These diagnostic sublineages were previously estimated as 33% and 18% in the Central Asian populations (31). We determined 13 (2.81%) individuals with these haplotypes in the TR population, therefore, their estimated contributions range from 0.0281/0.329×100=8.5% to 0.0281/0.180×100=15.6%. Central Asian specific mtDNA haplotypes were previously identified as D4c and G2a and their total frequency in the Central Asian populations was estimated as 8% (32). We determined 5 (0.65%) individuals with these haplotypes in the TR population therefore, their estimated contribution was 0.0065/0.08×100= 8.13%

4. Variome characterization

4.1. Derived allele frequencies (DAFs): Ancestral sequences for *Homo sapiens* (GRCh37), which were generated using the information from ENSEMBL compara and include the multiple sequence alignment of six primates, were downloaded from the 1000GP FTP site. WES and WGS VCF files were separately annotated with the ancestral alleles using Jvarkit, vcfancestralalleles tool (33). gnomAD WES and WGS VCFs and GME variants were downloaded and annotated using the same ancestral alleles. DAFs were calculated only for variant sites where an ancestral allele is present.

4.2. Functional annotation: Variants were annotated by ENSEMBL v.87 using SnpEff v.4.4 to determine variant functional region and impact on the assigned gene (34, 35). The high-confidence predicted loss of function variants (HC-pLoFs) (frameshift, essential splice site, stop gain, stop loss and start loss) were detected using LOFTEE, which is a VEP plugin designed to identify HC-LoF variants based on their ancestral state, transcript information, and splice prediction (36). Thus, the following low-confidence LoF variants flagged by LOFTEE were filtered out: variants for which the purported LoF allele is the ancestral state (across primates); stop-gain and frameshift variants in the last 5% of the transcript, or in an exon with noncanonical splice sites around it (i.e., intron does not start with GT and end with AG); and splice site variants in small introns (<15 bp), in an intron with a noncanonical splice site or rescued by nearby in-frame splice sites. We also performed transcript expression-aware annotation for single nucleotide pLoF variants. Proportion expression across transcripts (pext) is a measure of the proportion of the total transcriptional output from a gene that would be affected by the variant and can be used as a proxy for the functional importance of pLoF variants (37). The classification of the missense variants according to their predicted deleteriousness was performed using PolyPhen-2, SIFT, and CADD. PolyPhen-2 classifies the missense variants as B (benign), P (possibly damaging) or D (probably damaging) whereas SIFT classifies them as T (Tolerated) or D (deleterious) (38-40). We categorized the missense variants as deleterious if they were listed as “D” in both PolyPhen-2 and SIFT and had a CADD score > 20; it was classified as “other missense” in the rest of the outcomes. Variants were also annotated by the ANNOVAR v.2019Oct24 tool using the data from dbnsfp35a, which includes PolyPhen-2, SIFT, CADD, and GERP++ scores, gnomAD, 1000GP, the NHBLI GO Exome Sequencing Project (ESP) Exome variant server and GME databases (4, 12, 14, 41, 42). Annotations were performed separately for “high-quality” WES (n = 1,123,248) and WGS (n = 45,981,720) variants. Additionally, pLOF variants were annotated with gnomAD_pLI scores. Then, we listed variants detected both in WES and WGS (n = 365,489) and re-calculated their allele frequencies (AF). A variant was classified as “Novel”, if it had no record in dbSNP build 151, gnomAD, 1000GP, and ESP Exome variant server; it was deemed to be “Common”, if the variant had an AF $\geq$ 0.01 in any of the above-mentioned databases. If the variant had an AF < 0.01 in all databases, it was classified as “Rare” (Table S6).

4.3. Human knockouts: 1,294 homozygous HC-pLoF variants were identified in the TR Variome and 723 of these had an AF < 0.01. We removed homozygous HC-pLoFs variants,

which are found in phase with another variant, altering its functional interpretation. Additionally, homozygous pLOFs of gnomAD, and 1000GP were extracted and previously published lists of human-knockouts including Iceland, GME, PROMIS, British Pakistani, and GenomeAsia were downloaded and compared with our list of rare homozygous pLoF variants (6, 14, 43-45). The common homozygous pLoFs, which have a frequency in the TR population ≥ 0.01, were listed using HC-pLoF variants. The list was compared to the previously published list of common homozygous pLoFs in the ExAC and gnomAD (46). Among 38 technical replicates, three individuals carried three different homozygous rare pLOFs (p.Ser177fs in *PSG4,* p.Leu251fs in *FAM166A*, c.1259-1G>C in *ACOT9*). Thus, we validated these 3 pLoFs.

4.4. Clinically relevant variants: We annotated all variants that were identified in the TR Variome against Human Gene Mutation Database (HGMD) Professional v.2020.2 (47), ClinVar (Accessed September 9th 2020) (48), and Online Inheritance in Man (OMIM) (Accessed December 10th 2019). Only disease-causing pathological mutations (DMs) in HGMD and pathogenic or pathogenic/likely pathogenic variants in ClinVar were used for further analyses. Inheritance types of the phenotypes were extracted from the OMIM database, where applicable.

5. Per-genome variant summary and imputation panel

To compare the genetic structure of the TR population with other populations in terms of genome-wide variation, we first catalogued high-quality variants from the WGS dataset of the TR Variome with up to 20% missingness. We calculated the number of variant sites and singletons using BCFtools. The mean concordance rate between technical replicates of the WGS data ($n = 6$) for the singleton variants was 0.985 ± 0.003. We used a similar approach to that of previous publications for the generation of the TR reference panel and the evaluation of the imputation performance. Haplotypes of 773 TR individuals were constructed for each autosomal chromosome with BEAGLE v5.1 using the high-quality SNPs sequenced with WGS (49). The BEAGLE genotypes re-phased using SHAPEIT v2 to generate the final TR reference panel (50). Re-phasing was performed using default parameters except for a window size of 0.5, as it produces more accurate results for sequencing data. To evaluate the performance of the TR reference panel for predicting missing genotypes, we randomly subsampled 73 individuals by extracting their genotypes from unphased WGS data and removed their haplotypes from the TR reference panel. Then, using chromosome 20 variants from the 73 individuals, we generated a Pseudo-GWAS panel, which comprised the 44,367 SNPs represented on Infinium Omni2.5-8 Kit. 1000GP Phase 3 haplotypes were downloaded from IMPUTE2 website for comparison with the new TR reference panel (51). The imputation was performed by IMPUTE2 on chromosome 20 split into 5 Mb chunks with 250 kb buffer regions using: 1) 1000GP reference panel, 2) TR reference panel, 3) TR + 1000GP reference panels to predict the “masked” genotypes of the 73 individuals. The TR panel comprised only SNPs whereas the 1000GP contains SNPs, short indels plus copy number variations. We used the default parameters of IMPUTE2 except for setting k_hap (Number of reference haplotypes used as templates) to 10,000 since diverse reference panels could contain more useful haplotypes than expected. Squared Pearson’s correlation coefficients (R^2) were calculated between the masked sequence genotypes (0,1,2) and the imputed genotype dosages (0-2), to compare the performance of imputation using each reference haplotype panel. The R^2 results were plotted against non-overlapping AF bins. A Wilcoxon rank-sum test was performed to evaluate the statistical significance of the R^2 results. The summary file produced by IMPUTE2 was used to show the number of variants with different expected R^2 results and expected AF bins for each reference panel.

The phased WGS data of the Simons Genome Diversity Project (SGDP) were used to evaluate the performance of the TR reference panel for imputing genotypes of neighboring populations (52). We extracted chromosome 20 variants of the Balkan, Caucasus and Middle Eastern populations from the SGDP dataset; generated a pseudo-GWAS panel; imputed genotypes using 1000GP, TR, and TR+1000GP reference panels; and calculated the R2 between the masked sequence genotypes and the imputed genotype dosages. The results were plotted against non-overlapping AF bins.

Supplementary Figures

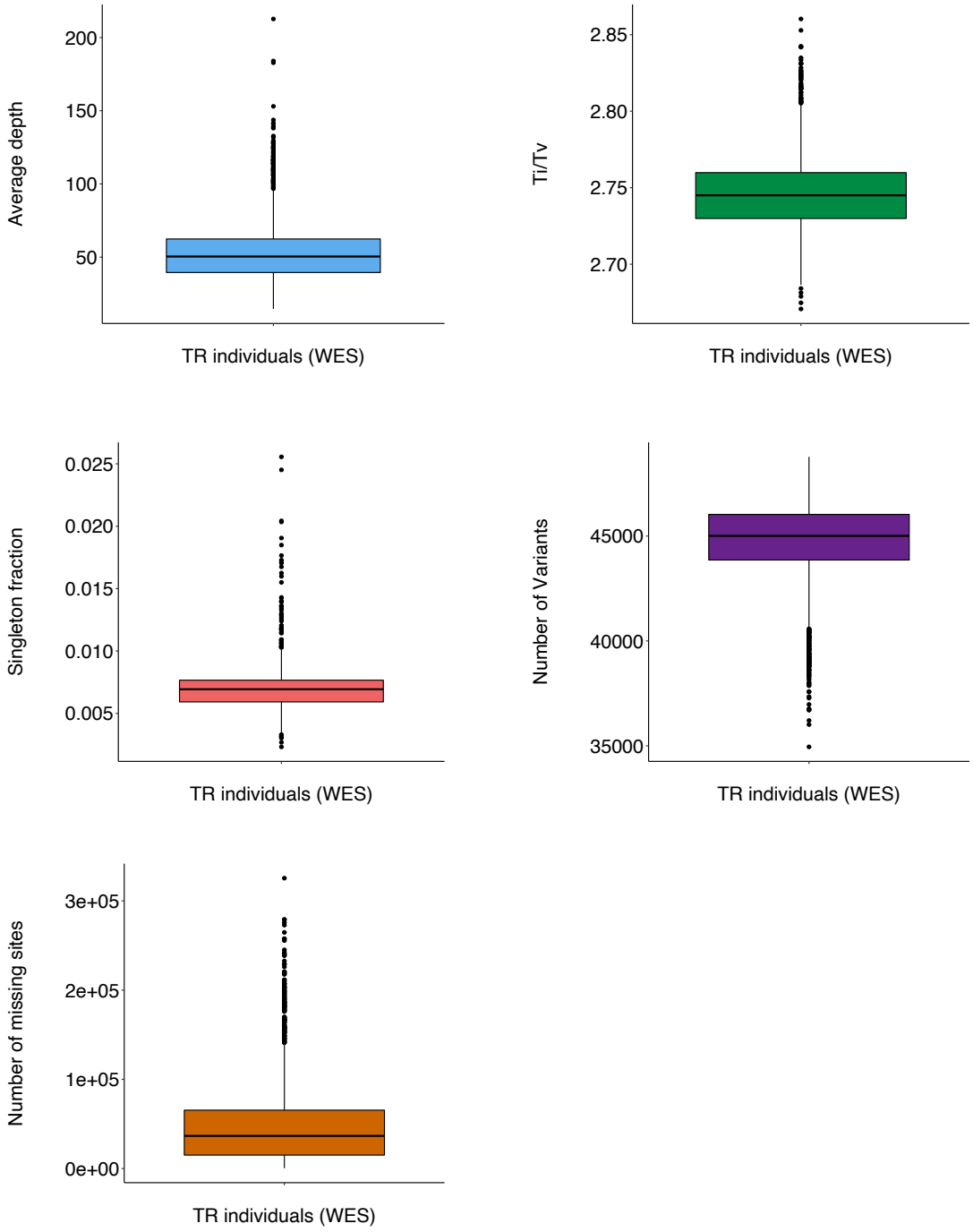

Fig. S1. Sample-based quality control metrics for TR-WES data. Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).

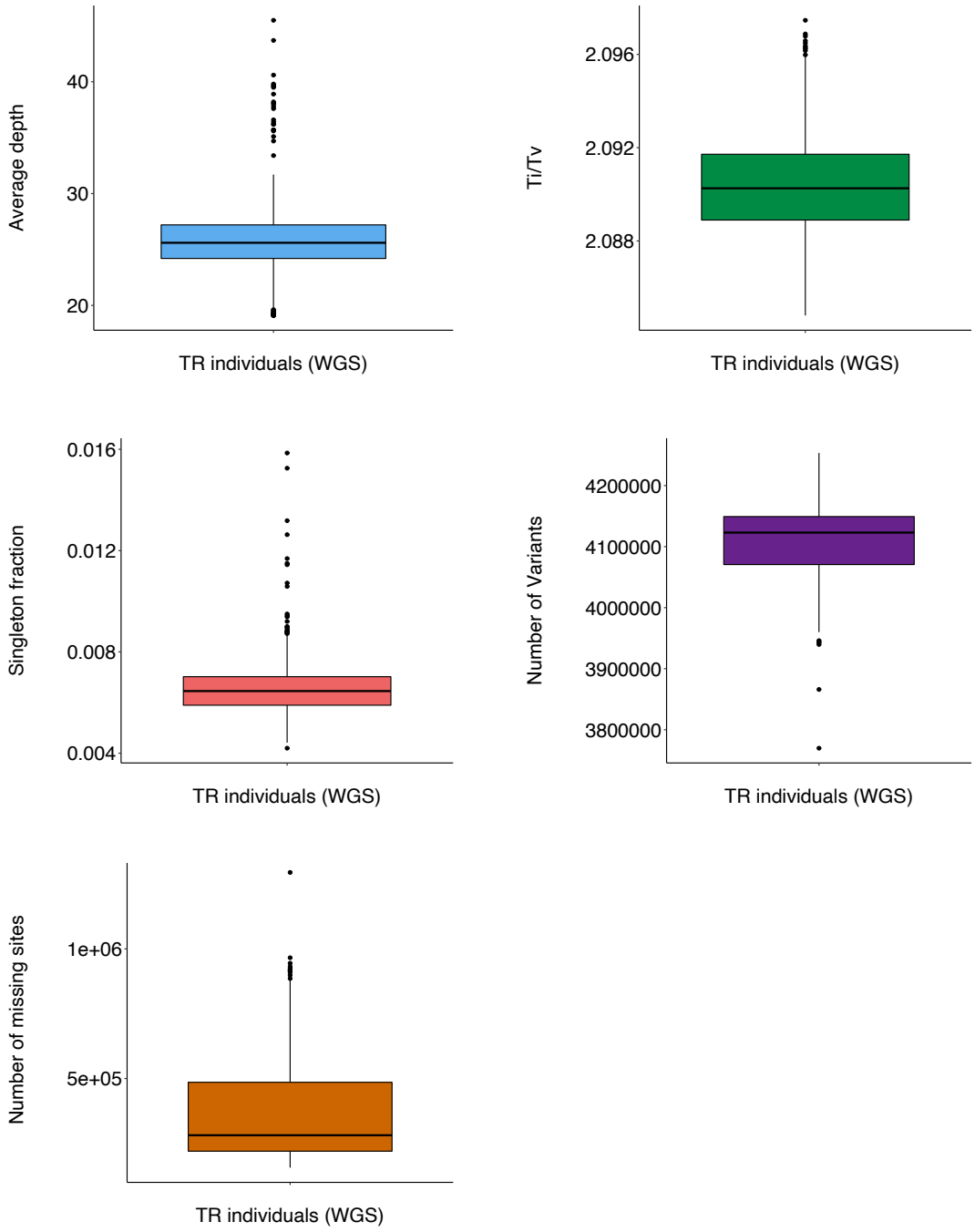

Fig. S2 Sample-based quality control metrics for TR-WGS data. Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).

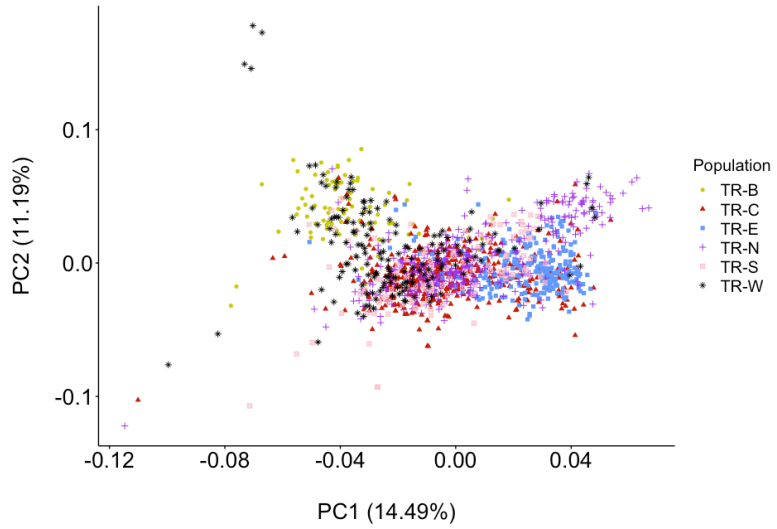

Fig. S3. Principal component analysis on TR individuals with known origin. Plots for the PC1 and PC2, which explain 14.49% and 11.19% of the total variation seen in Turkey (*n* = 1,460). PC1 distinguishes TR-W, TR-B and TR-E subregions.

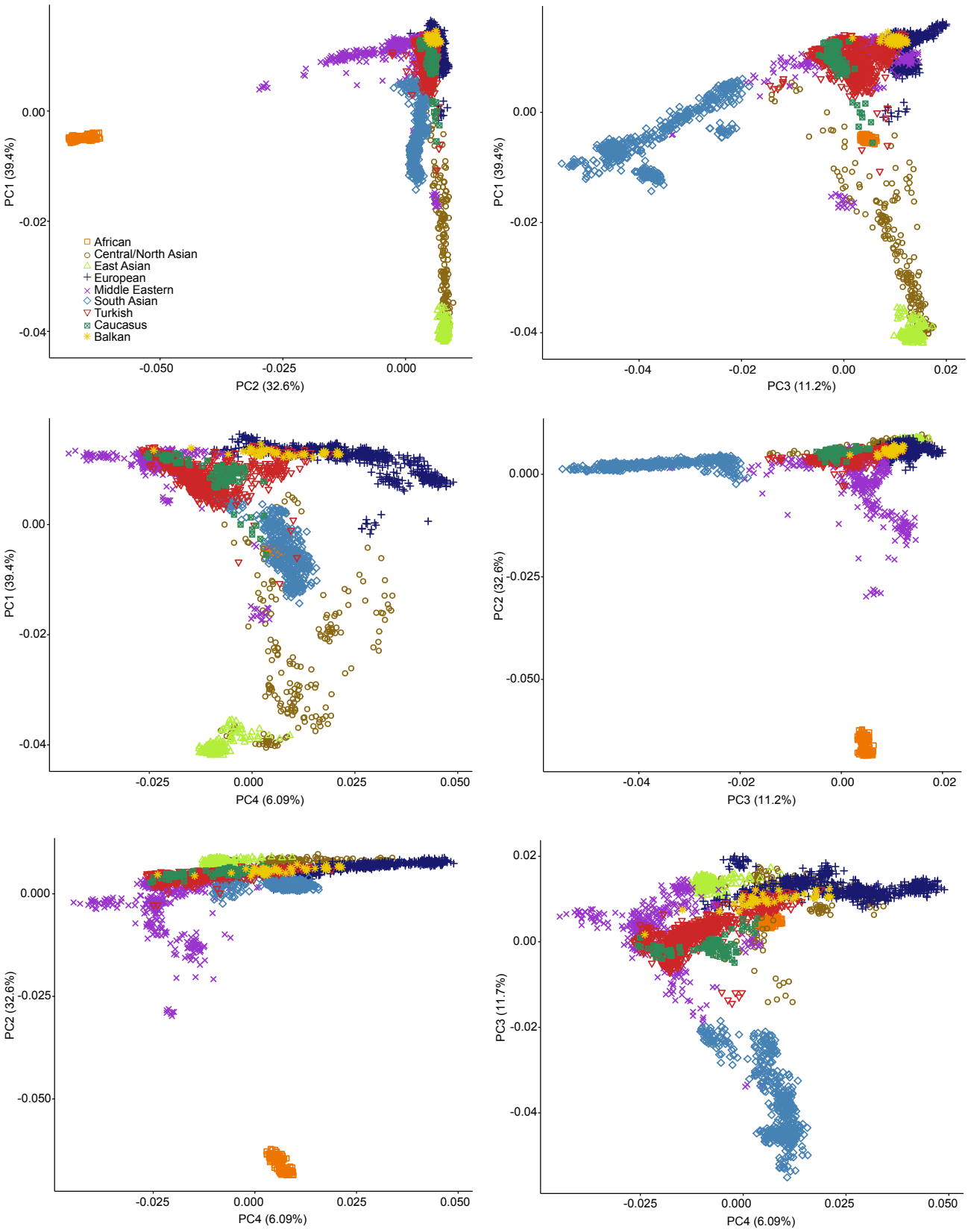
African
Central/North Asian
East Asian
European
Middle Eastern
South Asian
Turkish
Caucasus
Balkan
PC1 (39.4%)
PC2 (32.6%)
PC3 (11.2%)
PC4 (6.09%)
PC3 (11.7%)

Fig. S4. Principal-component analysis on the TR, Lazaridis et al. (13) and 1000GP populations. Plots for the first four principal components and percentages of variance explained. PC1 (39.4%) and PC2 (32.6%) separate East Asians, Central and North Asians, and Africans respectively from the other populations, while PC3 separates South Asians from the other populations. PC4 demonstrated the degree of variance between the Middle Eastern, Caucasus, TR, Balkan and European Populations.

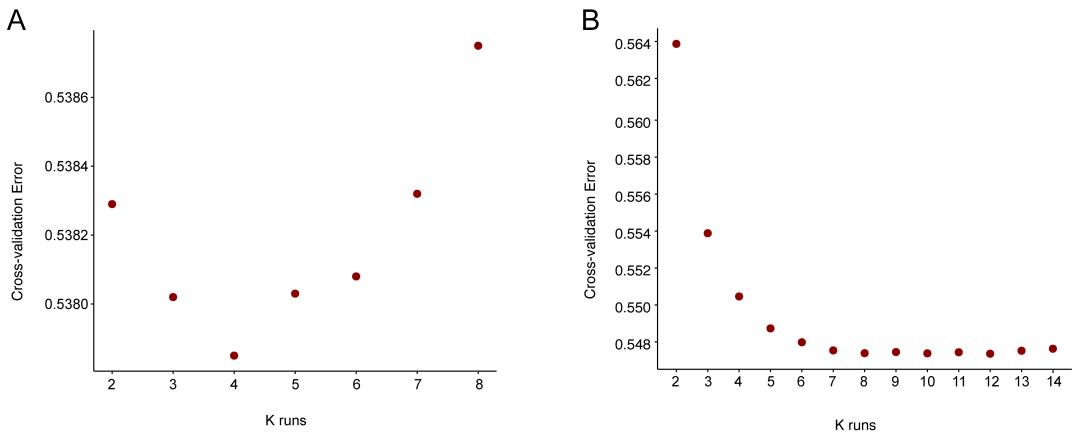

Fig. S5. ADMIXTURE cross-validation. (A) Cross-validation errors for TR subpopulations according to geographical regions of Turkey. *k* = 4 gave the lowest cross-validation error. (B) Cross-validation errors for the TR, Lazaridis et al. (13), and 1000GP samples. Analysis with eight ancestries (*k* = 12) resulted in the lowest cross-validation error.

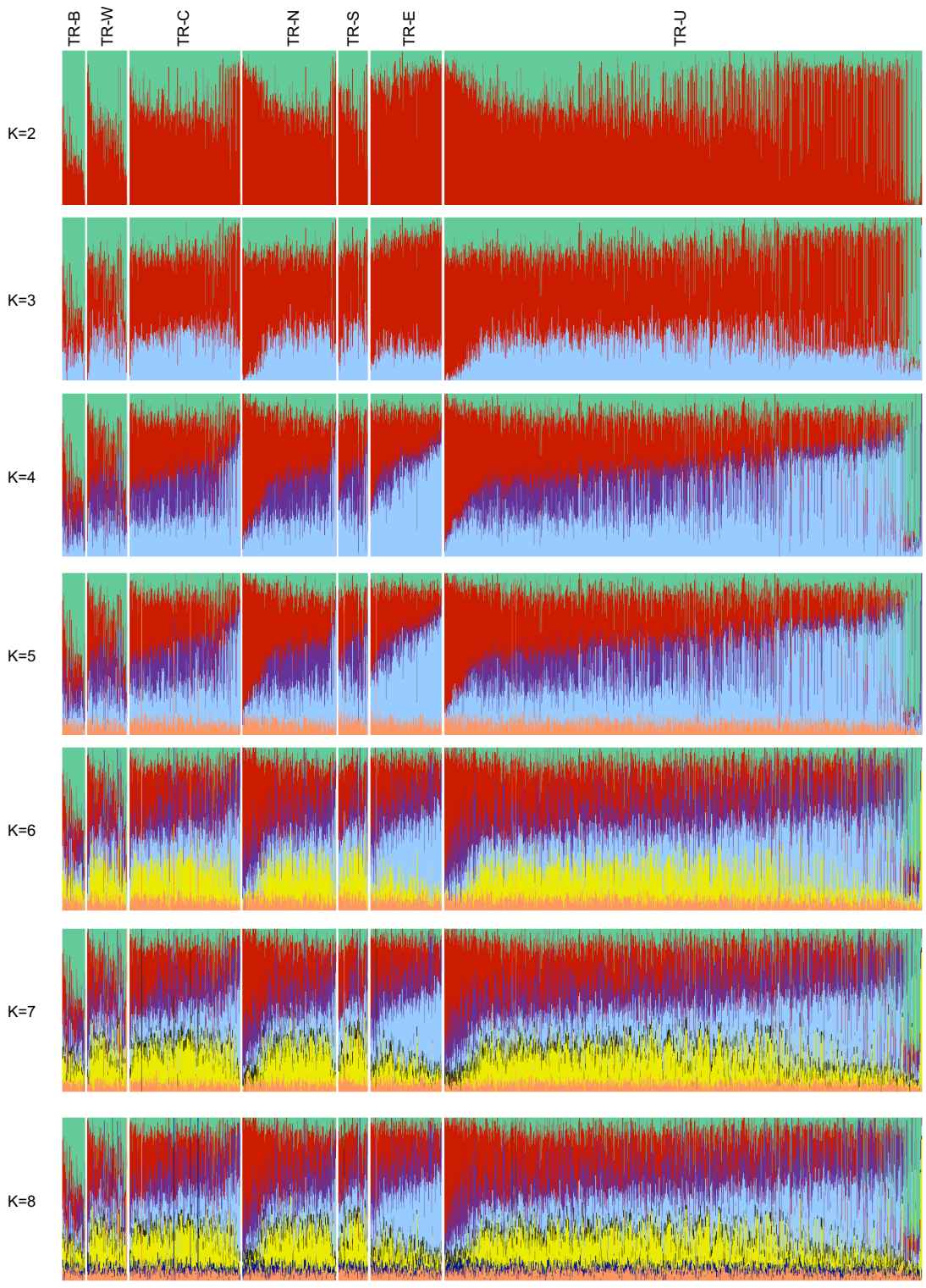

Fig. S6. Unsupervised ADMIXTURE analysis of the TR population for clusters *k* = 2 to *k* = 8. Samples (*n* = 3,362) grouped by geographical region and organized from west (left) to east (right). Each column represents an individual. The y axis represents a proportion ranging from 0 to 1. More color suggests multiple ancestral components.

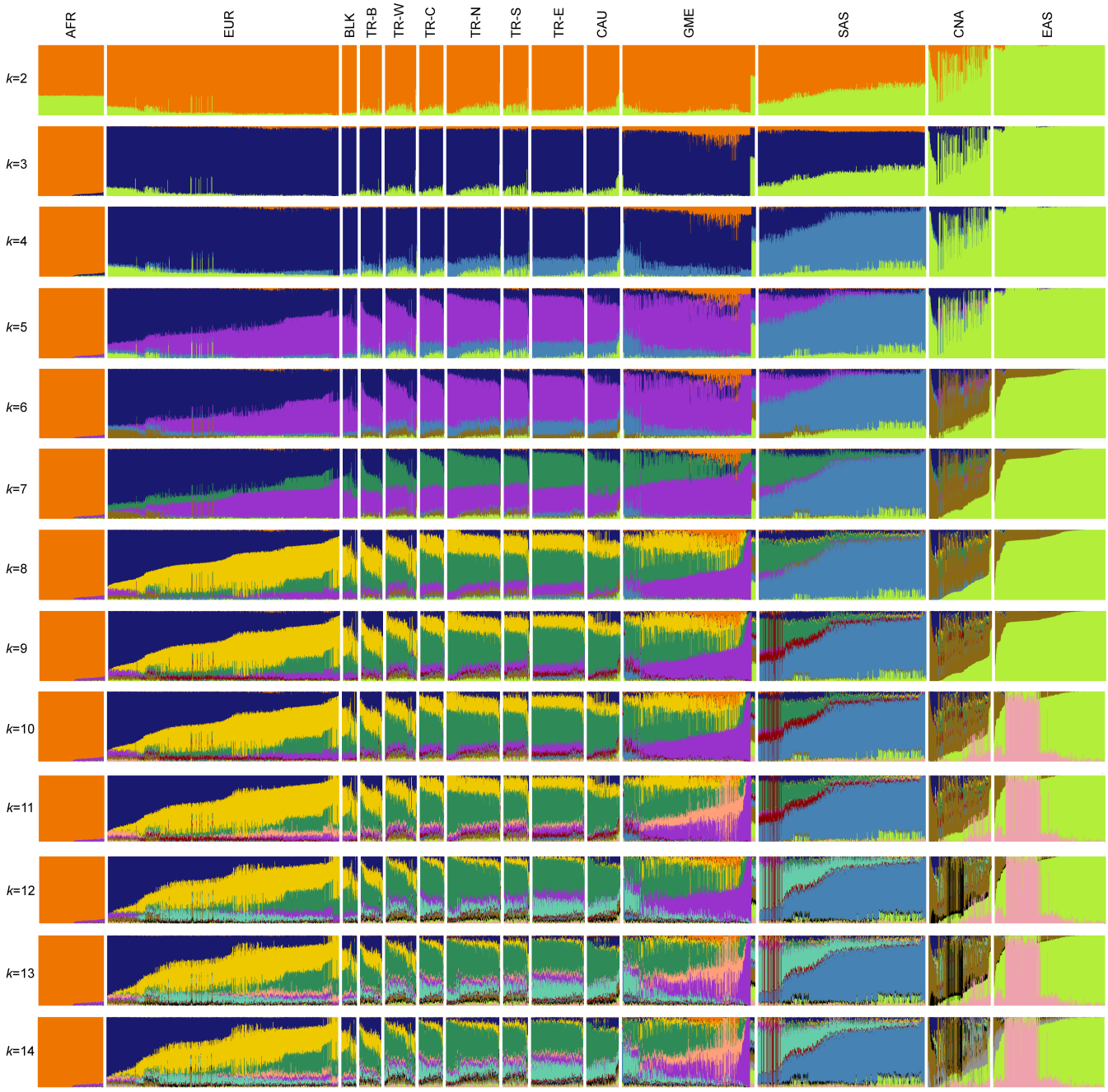

Fig. S7. Unsupervised ADMIXTURE analysis of the TR population in a global context for clusters *k* = 2 to *k* = 14. Samples from Turkey (*n* = 647), Lazaridis et al. (13) (*n* = 1,430) and 1000GP populations (*n* = 1,299) grouped by geographical region and organized from west (left) to east (right), showing trends of overlap.

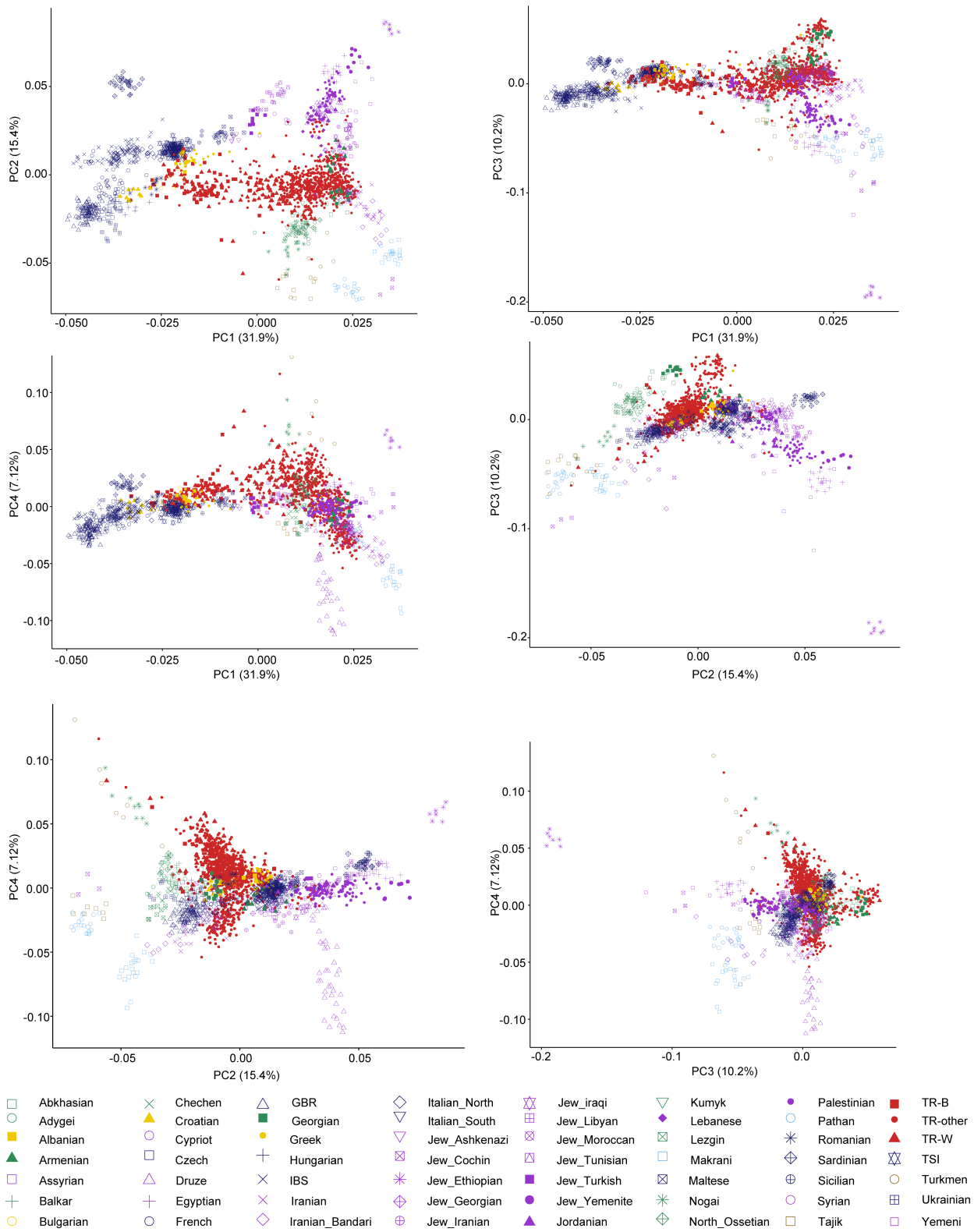

Fig. S8. Principal-component analysis on TR individuals and control populations in a regional context. Plots for the first four principal components and percentages of variance explained.

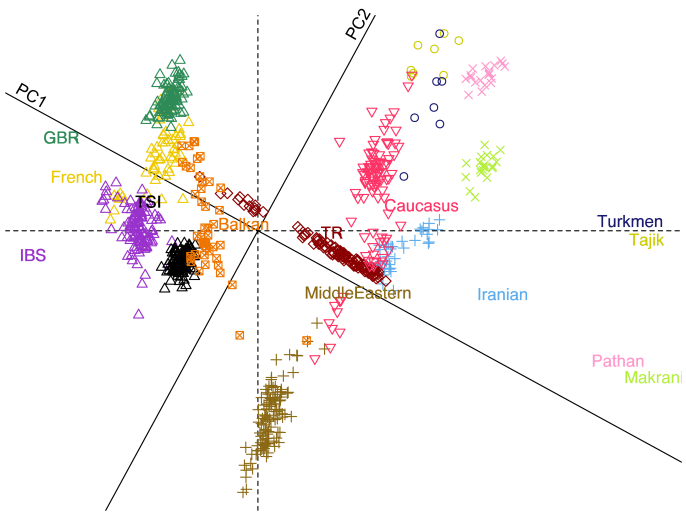

Fig. S9. Procrustes analysis on TR individuals and control populations. 100 individuals from TR were randomly selected and a Procrustes analysis was performed based on unprotected coordinates of geographical locations and PC1 and PC2 coordinates of TR,1000GP European, and populations from Lazaridis et al. (13).

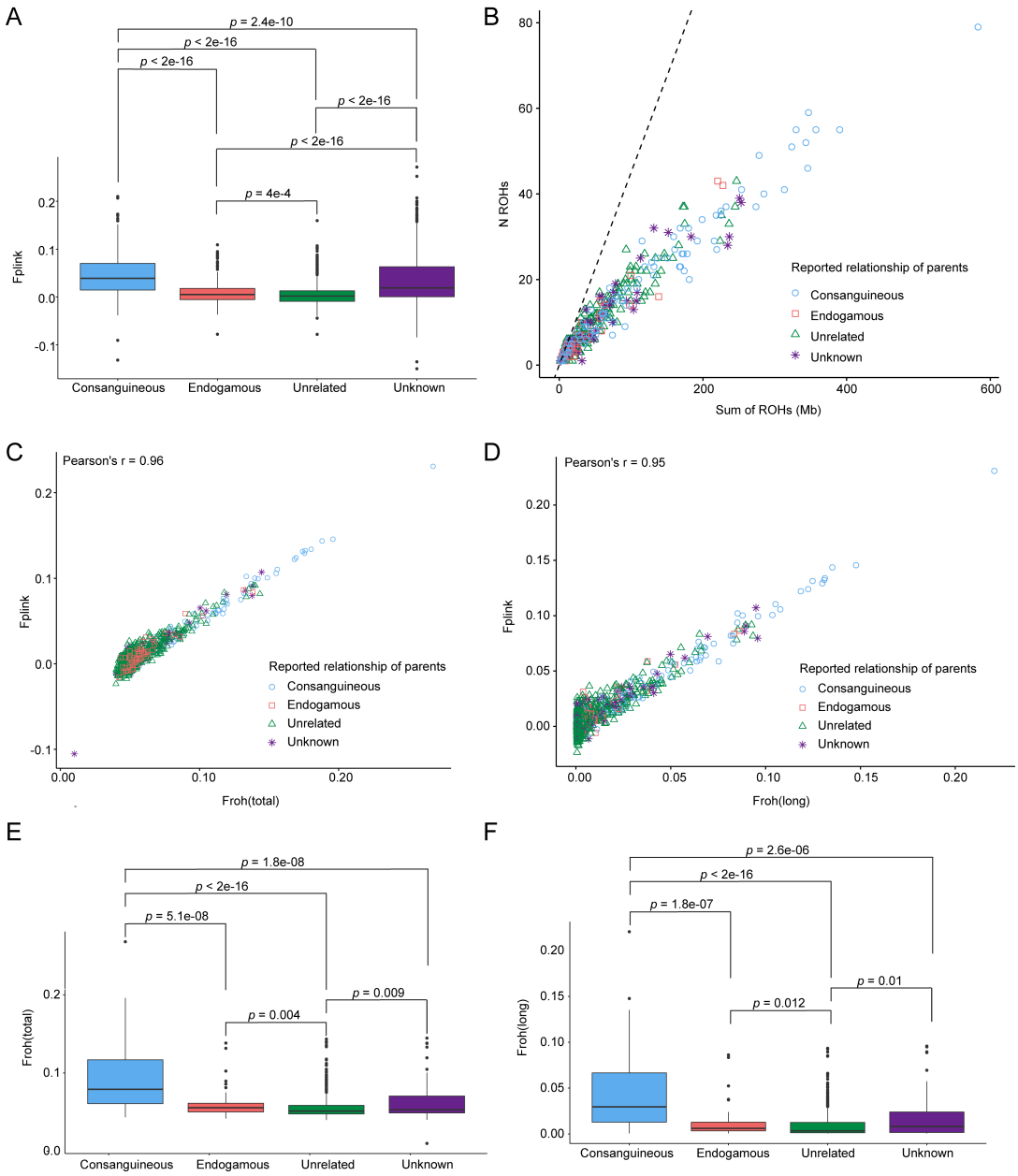

Fig. S10. Effect of consanguinity and endogamy on inbreeding coefficient, sum and number of ROHs. (A) Kruskal-Wallis test indicated that there was a statistically significant difference between the level of inbreeding coefficient (F_{plink}) with the degree of relationship of parents (H (3) = 557.46, P < 2.2e-16). (B) The correlation of F_{plink} and $F_{roh(total)}$, P < 2.2e-16. (C) The correlation of F_{plink} and $F_{roh(long)}$, P < 2.2e-16. (D) Kruskal-Wallis test indicated that there was a statistically significant difference between the level of $F_{roh(total)}$ with the degree of relationship of parents (H (3) = 112.2, P < 2.2e-16). (E) Kruskal-Wallis test indicated that there was a statistically significant difference between the level of F_{plink} with the degree of relationship of parents (H (3) = 97.135, P < 2.2e-16). (F) Number of ROHs compared to total length of ROHs. P values on the plots were obtained with a post-hoc Wilcoxon rank-sum test and adjusted with the Benjamini-Hochberg method. Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).

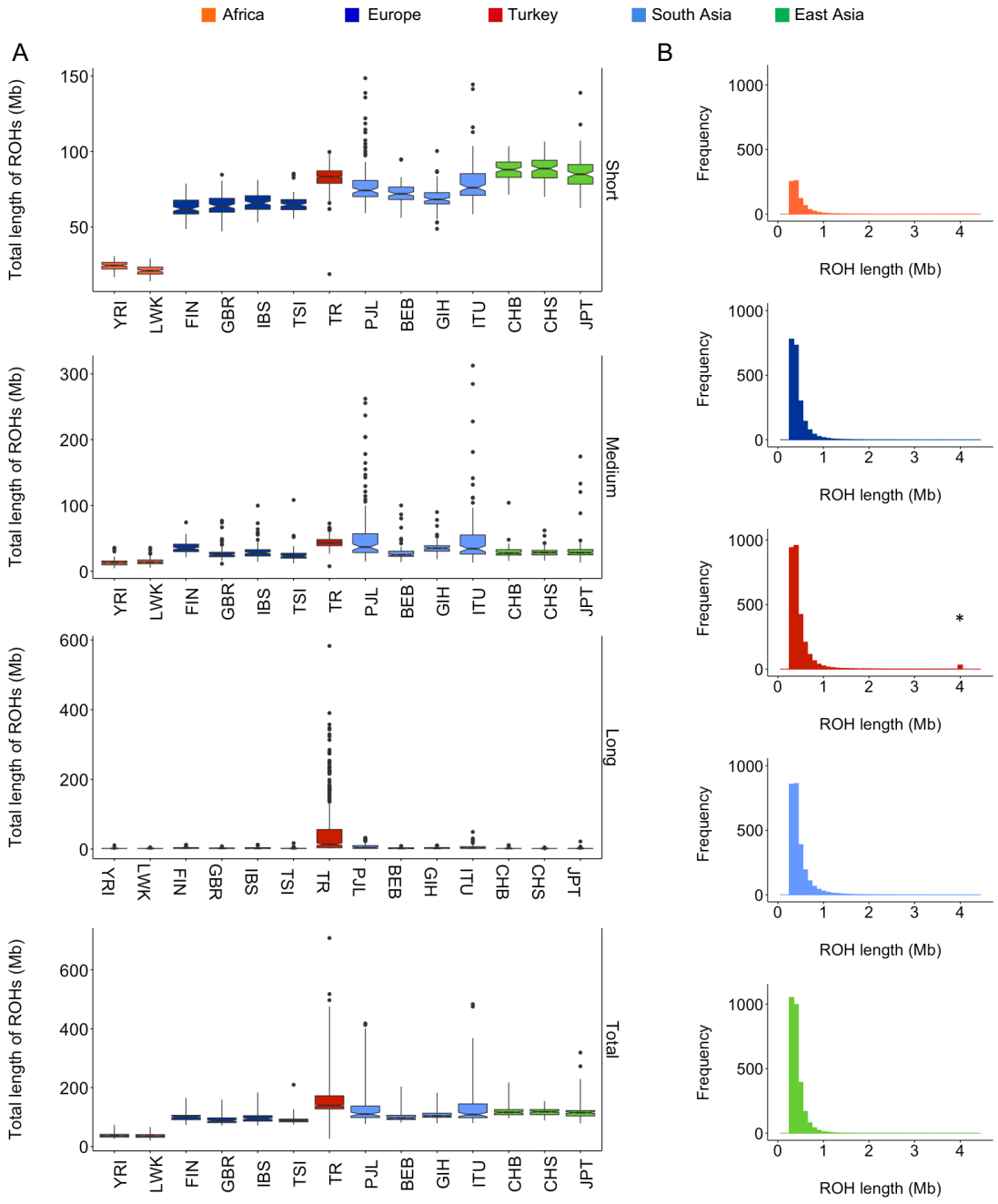

Fig. S11. Distributions of short, medium and long ROH correlate with patterns of bottlenecks and recent consanguinity. (A) Burden in samples of ROH grouped by length (short, < 516 Kb; medium 516-1,606 kb; long > 1,606 kb). The TR samples (red) showed a significantly increased number of long ROH in comparison to other populations. Box plots show the median (horizontal line), 25th percentile (lower edge), 75th percentile (upper edge), and minimum and maximum observations (whiskers).(B) Histograms of the frequencies of long ROH in the TR, African, European, South Asian and East Asian populations. Frequencies were calculated by dividing the number of ROH by the population size. ROH > 4 Mb in length are binned together (an asterisk indicates a small peak seen in the TR population).

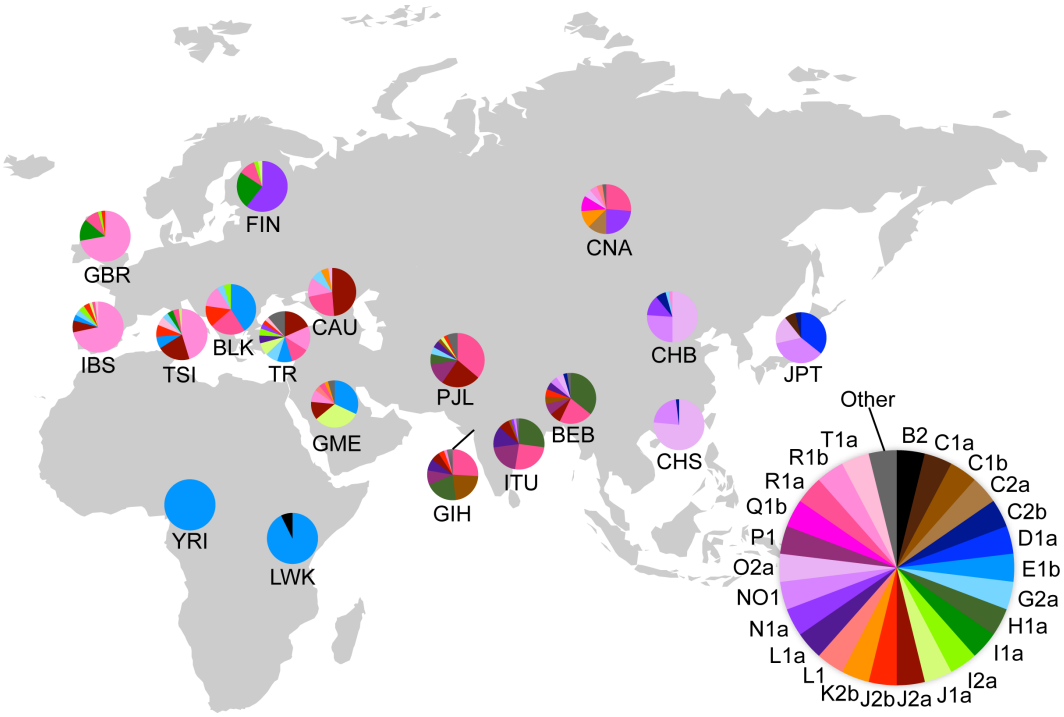

Fig. S12. Y-haplogroup distribution in the TR and control populations. The male samples from Albanian, Bulgarian, Croatian, and Greek populations were grouped together as Balkan (BLK); the male samples from Abkhazian, Adygei, Armenian, Balkar, Chechen, Georgian, Kumyk, Lezgin, Nogai, and North Ossetian populations were grouped together as Caucasus (CAU); and the male samples from Algerian, Assyrian, Bedouin, Cypriot, Druze, Egyptian, Hazara, Iranian, Jewish, Jordanian, Lebanese, Libyan, Moroccan, Mozabite, Palestinian, Saharawi, Saudi, Syrian, Tunisian, Yemeni populations were grouped together as Greater Middle Eastern (GME). Only main haplogroups are shown.

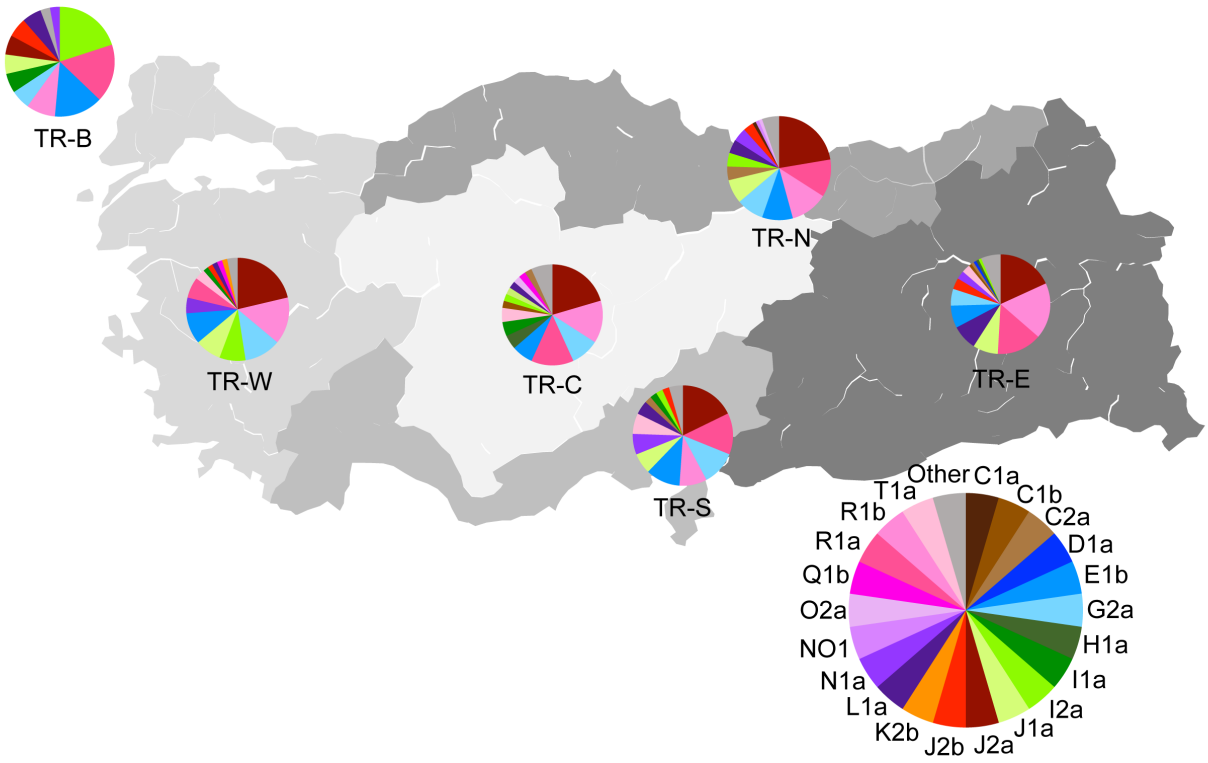

Fig. S13. Y-haplogroup distribution in the TR subregions. The pie charts demonstrate the Y chromosome haplogroups of TR males with known ancestral origin (*n* = 370). Only main haplogroups are shown.

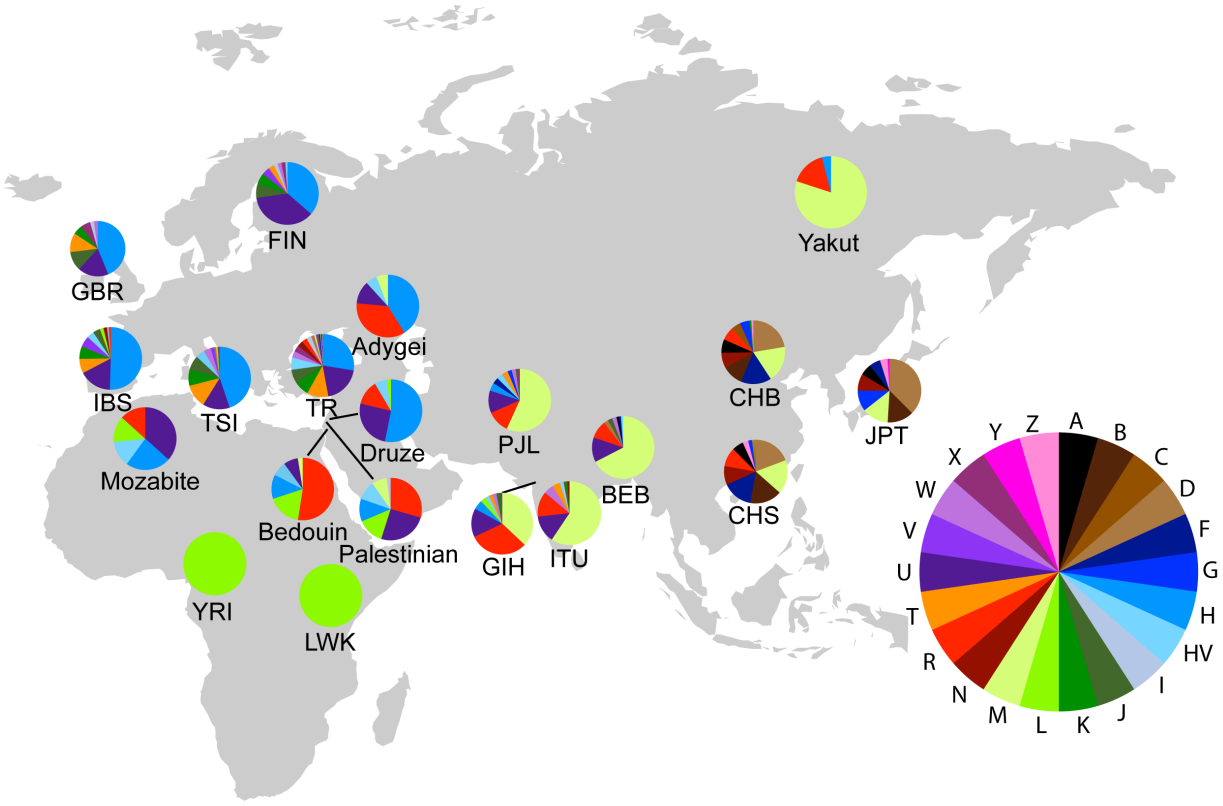

Fig. S14. mtDNA-haplogroup distribution in the TR and control populations. The mtDNA sequences of Adygei, Bedouin, Druze, Mozabite, Palestinian, and Yakut populations of the Human Genome Diversity Project (HGDP) were used in addition to the TR and 1000GP populations.

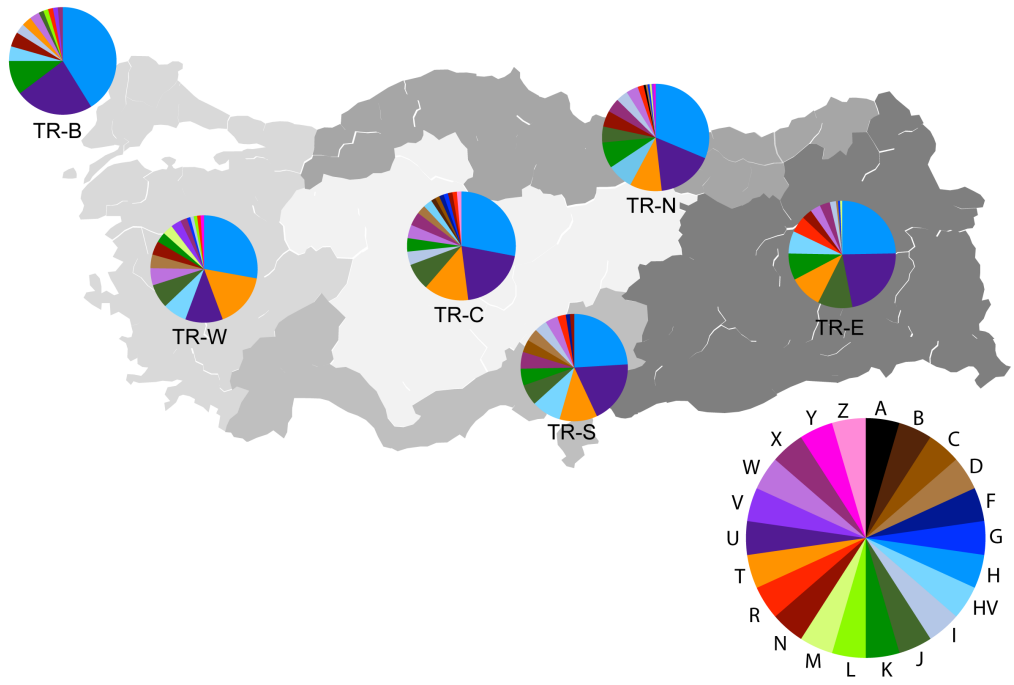

Fig. S15. mtDNA-haplogroup distribution in the TR subregions. The pie charts demonstrate the mt-DNA haplogroups of TR individuals with known ancestral origin (*n* = 647). Only main haplogroups are shown.

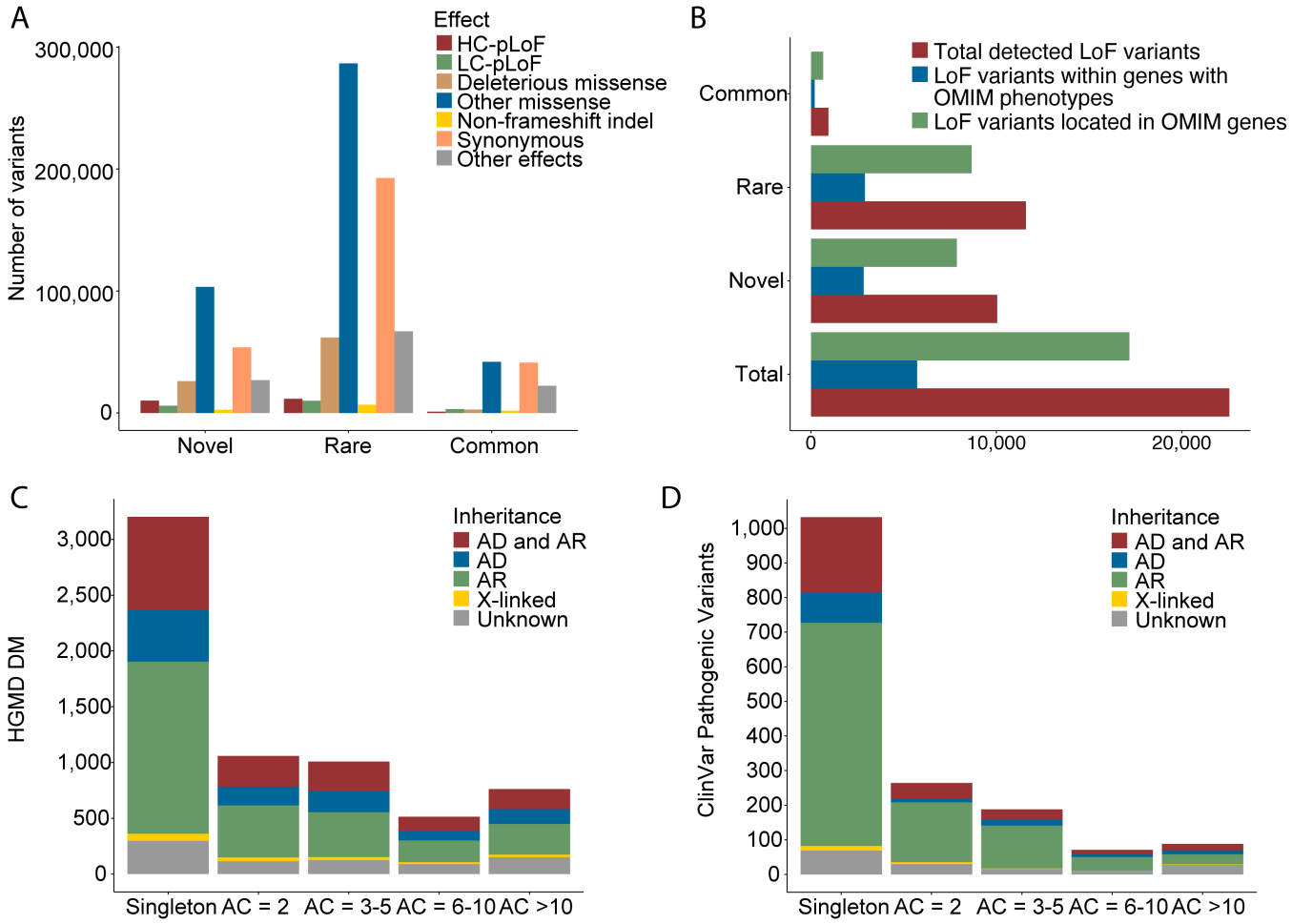

Fig. S16. Distribution of variants based on their frequency, pathogenicity and OMIM phenotypes. (A) Variants were classified as HC-pLoF, LC-pLoF, missense, non-frameshift indel, synonymous, and other effects as well as according to their frequency. Non-coding variants are not included in the Figure. (B) Proportions of HC-pLoF variants, which were grouped based on their frequency, their location on OMIM genes and the genes with OMIM phenotypes. (C) Distribution of disease-causing pathological mutations from HGMD and (D) pathogenic (P) or pathogenic/likely pathogenic (P/LP) variants from the ClinVar database, categorized based on their frequency and inheritance type as autosomal-dominant (AD) or autosomal-recessive (AR), X-linked or unknown.

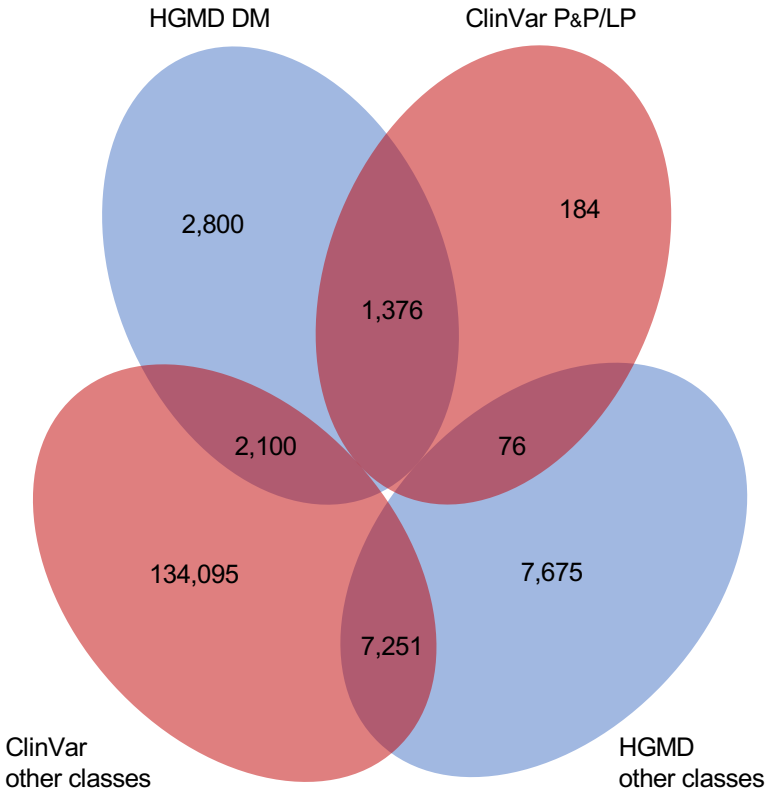

Fig. S17. The Venn Diagram showing the number of the TR variants catalogued in HGMD and/or ClinVar. HGMD had a higher number of records in the class of disease-causing pathological mutations (DM) when compared to the number of variants that were classified as pathogenic (P) or pathogenic/likely pathogenic (P/LP) in ClinVar.

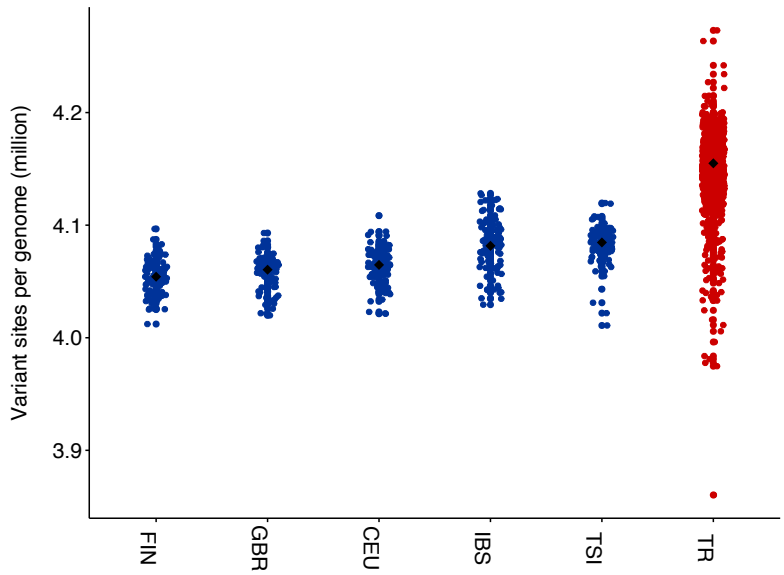

Fig. S18. The number of variant sites per genome for the 1000GP and TR populations. The average number of variant sites per genome is higher in the TR population than in the European populations.

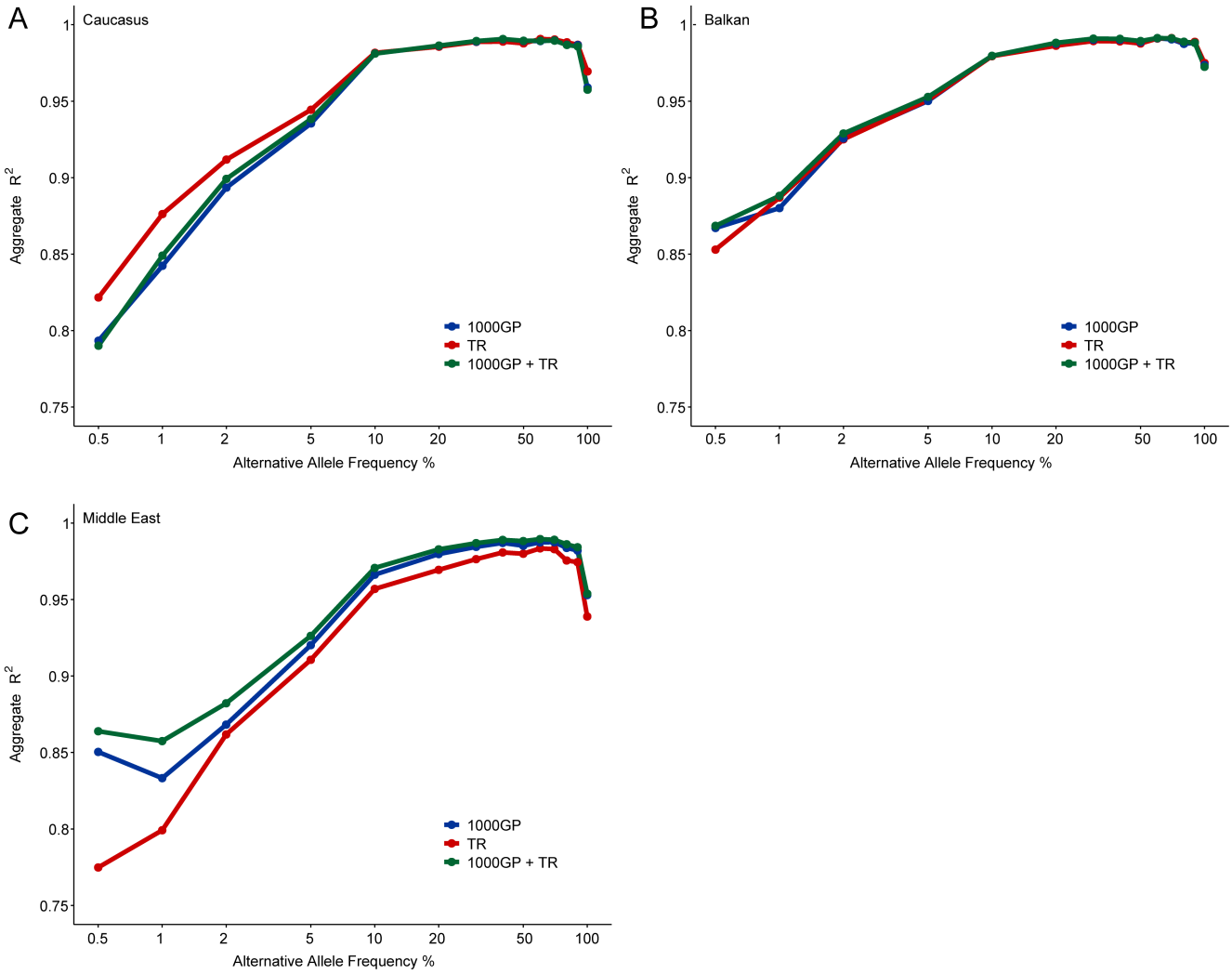

Fig. S19. Imputation accuracy of neighboring populations. Evaluation of imputation performance on chromosome 20. The aggregate squared Pearson correlation coefficient (R^2) was calculated for genotypes called from WGS and imputed genotypes and plotted against alternative allele frequency for the three reference haplotype panels. (A) Imputation accuracy for the Caucasus population (n = 13). Two-tailed Wilcoxon rank-sum test was used to assess the significance of the R^2 difference: P = 0.041 for the TR (mean ± s.d.: 0.96 ± 0.05) versus 1000GP (mean ± s.d.: 0.95 ± 0.06); P > 0.05 for the TR + 1000GP (mean ± s.d.: 0.95 ± 0.06) versus 1000GP. (B) Imputation accuracy for the Balkan population (n = 6). The significance of the R^2 difference: P > 0.05 for the TR (mean ± s.d.: 0.96 ± 0.04) versus 1000GP (mean ± s.d.: 0.96 ± 0.04); P = 0.009 for the TR + 1000GP (mean ± s.d.: 0.96 ± 0.04) versus 1000GP. (C) Imputation accuracy for the Middle Eastern population (n = 19). The significance of the R^2 difference: P = 0.009 for the TR (mean ± s.d.: 0.93 ± 0.07) versus 1000GP (mean ± s.d.: 0.95 ± 0.06); P = 0.009 for the TR + 1000GP (mean ± s.d.: 0.95 ± 0.05) versus 1000GP.

Supplementary Tables

Table S1. The TR Variome Summary

Cohort	*n*	Method
Amyotrophic lateral sclerosis	238	WES
Ataxia	269	WES
Delayed sleep phase disorder	19	WES
Essential tremors	154	WES
Obesity	765	WES
Parkinson's disease	53	WES
Polycystic ovarian syndrome	15	WES
Various neurological and immunological disorders	1,559	WES
Amyotrophic lateral sclerosis	792	WGS
Total	3,072 + 792 = 3,864	

Table S2. Demographics after exclusion of low quality and related samples

Cohort	Mean age	Gender (Female/Male)	Affected	Unaffected	Unknown
Amyotrophic lateral sclerosis	54.71 ± 14.77	395/514	698	211	0
Ataxia	44.3 ± 14.82	75/73	101	47	0
Delayed sleep phase disorder	28.38 ± 6.99	8/10	18	0	0
Essential tremors	52.37 ± 19.87	44/35	57	22	0
Obesity	38.87 ± 12.39	491/181	560	112	0
Parkinson's disease	47.18 ± 19.43	13/16	19	10	0
Polycystic ovarian syndrome	29 ± 10.11	9/0	9	0	0
Various neurological and immunological disorders	42.12 ± 17.88	673/825	32	26	1,440
Total	47.41± 16.41	1,708/1,654	1,494	428	1,440

Table S3. Sample based-quality control measures for the integration of the coding regions of WES and WGS data

Sequencing type	Whole genome		Whole exome	
Sample size	773		2,589	
Minimum variant depth	8		8	
Minimum allele count	1		1	
Mean variant depth	25		54	
Novel SNPs (% not in dbSNP 151)	0.29%		0.27%	
Transition/transversion	2.45		2.74	
	All	Novel	All	Novel
Number of variants	66,847	317	42,285	164
Heterozygotes	41,414	314	25,547	159
Variant homozygotes	22,433	3	16,738	5

Table S4. Populations included in the study

Population		Super population		Sequencing method	*n*	Study
YRI	Yoruba in Ibadan, Nigeria	AFR	African	WGS	108	1000GP
LWK	Luhya in Webuye, Kenya	AFR	African	WGS	96	1000GP
GWD	Gambian in Western Divisions in the Gambia	AFR	African	WGS	113	1000GP
MSL	Mende in Sierra Leone	AFR	African	WGS	85	1000GP
ESN	Esan in Nigeria	AFR	African	WGS	99	1000GP
ASW	Americans of African Ancestry in SW USA	AFR	African	WGS	61	1000GP
ACB	African Caribbeans in Barbados	AFR	African	WGS	96	1000GP
MXL	Mexican Ancestry from Los Angeles, CA, USA	AMR	American	WGS	64	1000GP
PUR	Puerto Ricans from Puerto Rico	AMR	American	WGS	104	1000GP
CLM	Colombians from Medellin, Colombia	AMR	American	WGS	94	1000GP
PEL	Peruvians from Lima, Peru	AMR	American	WGS	85	1000GP
-	Albanian	BLK	Balkan	H.O. array*/WGS	6/1	(13) /SGDP
-	Bulgarian	BLK	Balkan	H.O. array*/WGS	10/1	(13)/SGDP
-	Crete	BLK	Balkan	WGS	2	SGDP
-	Croatian	BLK	Balkan	H.O. array*	10	(13)
-	Greek	BLK	Balkan	H.O. array*/WGS	20/2	(13)/SGDP
-	Abkhazian	CAU	Caucasus	H.O. array*/WGS	9/2	(13)/SGDP
-	Adygei	CAU	Caucasus	H.O. array*/WGS/ Illumina HuHap 650k	17/2/17	(13)/SGDP/HGDP
-	Armenian	CAU	Caucasus	H.O. array*/WGS	10/2	(13)/SGDP
-	Balkar	CAU	Caucasus	H.O. array*	10	(13)
-	Chechen	CAU	Caucasus	H.O. array*/WGS	9/1	(13)/SGDP
-	Georgian	CAU	Caucasus	H.O. array*/WGS	10/2	(13)/SGDP
-	Kumyk	CAU	Caucasus	H.O. array*	8	(13)
-	Lezgin	CAU	Caucasus	H.O. array*/WGS	9/2	(13)/SGDP
-	Nogai	CAU	Caucasus	H.O. array*	9	(13)
-	North_Ossetian	CAU	Caucasus	H.O. array*/WGS	10/2	(13)/SGDP
-	Altaian	CNA	Central and North Asian	H.O. array*	7	(13)
-	Dolgan	CNA	Central and North Asian	H.O. array*	3	(13)
-	Even	CNA	Central and North Asian	H.O. array*	10	(13)
-	Kalmyk	CNA	Central and North Asian	H.O. array*	10	(13)
-	Kyrgyz	CNA	Central and North Asian	H.O. array*	9	(13)
-	Mansi	CNA	Central and North Asian	H.O. array*	8	(13)

-	Mongola	CNA	Central and North Asian	H.O. array*	6	(13)
-	Selkup	CNA	Central and North Asian	H.O. array*	10	(13)
-	Tajik	CNA	Central and North Asian	H.O. array*	8	(13)
-	Tubalar	CNA	Central and North Asian	H.O. array*	22	(13)
-	Turkmen	CNA	Central and North Asian	H.O. array*	7	(13)
-	Tuvinian	CNA	Central and North Asian	H.O. array*	10	(13)
-	Ulchi	CNA	Central and North Asian	H.O. array*	25	(13)
-	Uygur	CNA	Central and North Asian	H.O. array*	10	(13)
-	Uzbek	CNA	Central and North Asian	H.O. array*	10	(13)
-	Yakut	CNA	Central and North Asian	H.O. array*/ Illumina HuHap 650k	20/25	(13)/HGDP
-	Yukagir	CNA	Central and North Asian	H.O. array*	19	(13)
CHB	Han Chinese in Beijing, China	EAS	East Asian	WGS	103	1000GP
JPT	Japanese in Tokyo, Japan	EAS	East Asian	WGS	104	1000GP
CHS	Southern Han Chinese	EAS	East Asian	WGS	104	1000GP
CDX	Chinese Dai in Xishuangbanna, China	EAS	East Asian	WGS	93	1000GP
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS	East Asian	WGS	99	1000GP
-	Daur	EAS	East Asian	H.O. array*	9	(13)
-	Oroqen	EAS	East Asian	H.O. array*	9	(13)
-	Tu	EAS	East Asian	H.O. array*	10	(13)
-	Xibo	EAS	East Asian	H.O. array*	7	(13)
-	Chuvash	EUR	European	H.O. array*	10	(13)
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR	European	WGS	99	1000GP
TSI	Toscani in Italia	EUR	European	WGS	107	1000GP
FIN	Finnish in Finland	EUR	European	WGS	99	1000GP
GBR	British in England and Scotland	EUR	European	WGS	89	1000GP
IBS	Iberian population in Spain	EUR	European	WGS	107	1000GP
-	Basque	EUR	European	H.O. array*	29	(13)
-	Belarusian	EUR	European	H.O. array*	10	(13)
-	Czech	EUR	European	H.O. array*	10	(13)
-	Estonian	EUR	European	H.O. array*	10	(13)
-	French	EUR	European	H.O. array*	61	(13)
-	Hungarian	EUR	European	H.O. array*	20	(13)

-	Icelandic	EUR	European	H.O. array*	12	(13)
-	italian_North	EUR	European	H.O. array*	20	(13)
-	Italian_South	EUR	European	H.O. array*	6	(13)
-	Lithuanian	EUR	European	H.O. array*	10	(13)
-	Maltese	EUR	European	H.O. array*	8	(13)
-	Mordovian	EUR	European	H.O. array*	10	(13)
-	Norwegian	EUR	European	H.O. array*	11	(13)
-	Orcadian	EUR	European	H.O. array*	13	(13)
-	Romanian	EUR	European	H.O. array*	10	(13)
-	Russian	EUR	European	H.O. array*	22	(13)
-	Saami	EUR	European	H.O. array*	1	(13)
-	Sardinian	EUR	European	H.O. array*	27	(13)
-	Sicillian	EUR	European	H.O. array*	11	(13)
-	Ukranian	EUR	European	H.O. array*	9	(13)
NEA	North East Africam	GME	Middle Eastern	WES**	368	(14)
NWA	North West African	GME	Middle Eastern	WES**	99	(14)
AP	Arabian Peninsula	GME	Middle Eastern	WES**	171	(14)
SD	Syrian Desert	GME	Middle Eastern	WES**	58	(14)
-	Algerian	GME	Middle Eastern	H.O. array*	7	(13)
-	Assyrian	GME	Middle Eastern	H.O. array*	11	(13)
-	Bedouin	GME	Middle Eastern	H.O. array*/WGS/ Illumina HuHap 650k	44/2/40	(13)/SGDP/HGDP
-	Cypriot	GME	Middle Eastern	H.O. array*	8	(13)
-	Druze	GME	Middle Eastern	H.O. array*/WGS/ Illumina HuHap 650k	39/2/47	(13)/SGDP/HGDP
-	Egyptian	GME	Middle Eastern	H.O. array*	18	(13)
-	Hazara	GME	Middle Eastern	H.O. array*	14	(13)
-	Iranian	GME	Middle Eastern	H.O. array*/WGS	38/2	(13)/SGDP
-	Iranian_Bandari	GME	Middle Eastern	H.O. array*	8	(13)
-	Jew_Ashkenazi	GME	Middle Eastern	H.O. array*	7	(13)
-	Jew_Cochin	GME	Middle Eastern	H.O. array*	5	(13)
-	Jew_Ethiopian	GME	Middle Eastern	H.O. array*	7	(13)
-	Jew_Georgian	GME	Middle Eastern	H.O. array*	7	(13)
-	Jew_Iranian	GME	Middle Eastern	H.O. array*	9	(13)
-	Jew_Iraqi	GME	Middle Eastern	H.O. array*/WGS	6/2	(13)/SGDP
-	Jew_Libyan	GME	Middle Eastern	H.O. array*	9	(13)
-	Jew_Moroccan	GME	Middle Eastern	H.O. array*	6	(13)
-	Jew_Tunisian	GME	Middle Eastern	H.O. array*	7	(13)
-	Jew_Turkish	GME	Middle Eastern	H.O. array*	8	(13)
-	Jew_Yemenite	GME	Middle Eastern	H.O. array*/WGS	8/2	(13)/SGDP
-	Jordanian	GME	Middle Eastern	H.O. array*/WGS	9/3	(13)/SGDP

-	Lebanese	GME	Middle Eastern	H.O. array*	28	(13)
-	Libyan	GME	Middle Eastern	H.O. array*	6	(13)
-	Moroccan	GME	Middle Eastern	H.O. array*	10	(13)
-	Mozabite	GME	Middle Eastern	H.O. array*/WGS/ Illumina HuHap 650k	21/2/30	(13)/SGDP/HGDP
-	Palestinian	GME	Middle Eastern	H.O. array*/WGS/ Illumina HuHap 650k	38/3/51	(13)/SGDP/HGDP
-	Samartian	GME	Middle Eastern	WGS	1	SGDP
-	Saharawi	GME	Middle Eastern	H.O. array*	6	(13)
-	Saudi	GME	Middle Eastern	H.O. array*	8	(13)
-	Syrian	GME	Middle Eastern	H.O. array*	8	(13)
-	Tunisian	GME	Middle Eastern	H.O. array*	8	(13)
-	Yemeni	GME	Middle Eastern	H.O. array*	6	(13)
GIH	Gujarati Indian from Houston, Texas, USA	SAS	South Asian	WGS	100	1000GP
PJL	Punjabi from Lahore, Pakistan	SAS	South Asian	WGS	95	1000GP
BEB	Bengali from Bangladesh	SAS	South Asian	WGS	86	1000GP
STU	Sri Lankan Tamil from the UK	SAS	South Asian	WGS	102	1000GP
ITU	Indian Telugu from the UK	SAS	South Asian	WGS	101	1000GP
-	Balochi	SAS	South Asian	H.O. array*	20	(13)
-	Brahui	SAS	South Asian	H.O. array*	21	(13)
-	Burusho	SAS	South Asian	H.O. array*	23	(13)
-	Kalash	SAS	South Asian	H.O. array*	18	(13)
-	Makrani	SAS	South Asian	H.O. array*	20	(13)
-	Pathan	SAS	South Asian	H.O. array*	19	(13)
-	Sindhi	SAS	South Asian	H.O. array*	18	(13)
TR-B	Turkish with Balkan Ancestry	TR	Turkish	WGS/WES	68/22	Current study
TR-W	Western Turkish	TR	Turkish	WGS/WES	97/60	Current study
TR-C	Central Turkish	TR	Turkish	WGS/WES	75/366	Current study
TR-N	Northern Turkish	TR	Turkish	WGS/WES	166/206	Current study
TR-S	Southern Turkish	TR	Turkish	WGS/WES	79/37	Current study
TR-E	Eastern Turkish	TR	Turkish	WGS/WES	162/122	Current study
TR-U	Turkish with unknown origin	TR	Turkish	WGS/WES	126/1,776	Current study
-	Turkish	TR	Turkish	H.O. array*	56	(13)

**: Affymetrix Human Origins array
**: Included only in Treemix analysis using allele frequency data of (14)

Table S5. F_{ST} for the TR subregions

	TR-B	TR-W	TR-C	TR-N	TR-S	TR-E
TR-B	-	0.001	0.002	0.002	0.002	0.003
TR-W	0.001	-	0.001	0.001	0	0.002
TR-C	0.002	0.001	-	0	0	0
TR-N	0.002	0.001	0	-	0	0.001
TR-S	0.002	0	0	0	-	0.001
TR-E	0.003	0.002	0	0.001	0.001	-
TR-U	0.001	0	0	0	0	0.001
Abkhasian	0.006	0.004	0.003	0.003	0.004	0.004
Adygei	0.006	0.005	0.005	0.005	0.005	0.005
Albanian	0	0.002	0.003	0.003	0.003	0.005
Armenian	0.004	0.002	0.001	0.001	0.001	0.001
Assyrian	0.007	0.005	0.004	0.004	0.004	0.003
Balkar	0.005	0.004	0.003	0.003	0.004	0.004
Bulgarian	0	0.002	0.003	0.004	0.003	0.005
Chechen	0.007	0.006	0.006	0.006	0.006	0.006
Croatian	0.003	0.005	0.007	0.008	0.007	0.009
Cypriot	0.004	0.003	0.003	0.003	0.003	0.004
Czech	0.003	0.005	0.007	0.008	0.008	0.009
Druze	0.009	0.008	0.007	0.007	0.007	0.007
Egyptian	0.009	0.007	0.006	0.007	0.006	0.007
French	0.003	0.005	0.007	0.007	0.007	0.009
GBR	0.004	0.006	0.008	0.009	0.008	0.01
Georgian	0.007	0.005	0.004	0.003	0.004	0.004
Greek	0.001	0.002	0.003	0.003	0.003	0.004
Hungarian	0.002	0.004	0.006	0.007	0.006	0.008
IBS	0.003	0.004	0.006	0.007	0.006	0.008
Iranian	0.005	0.003	0.002	0.002	0.002	0.001
Iranian_Bandari	0.01	0.007	0.006	0.007	0.006	0.006
Italian_North	0.002	0.003	0.004	0.005	0.004	0.006
Italian_South	0.014	0.014	0.015	0.015	0.015	0.016
Jew_Ashkenazi	0.005	0.005	0.006	0.006	0.006	0.007
Jew_Cochin	0.019	0.016	0.016	0.017	0.016	0.016
Jew_Ethiopian	0.036	0.033	0.033	0.034	0.031	0.033
Jew_Georgian	0.011	0.009	0.008	0.008	0.008	0.008
Jew_Iranian	0.01	0.008	0.007	0.008	0.007	0.007
Jew_iraqi	0.009	0.007	0.006	0.006	0.006	0.006
Jew_Libyan	0.012	0.011	0.011	0.011	0.011	0.012
Jew_Moroccan	0.007	0.006	0.007	0.007	0.006	0.007
Jew_Tunisian	0.012	0.011	0.011	0.011	0.011	0.011
Jew_Turkish	0.004	0.004	0.004	0.004	0.003	0.005
Jew_Yemenite	0.017	0.015	0.014	0.014	0.013	0.014
Jordanian	0.006	0.004	0.004	0.004	0.003	0.004
Kumyk	0.003	0.002	0.002	0.002	0.002	0.002

Lebanese	0.004	0.003	0.002	0.002	0.002	0.002
Lezgin	0.005	0.004	0.004	0.004	0.004	0.004
Makrani	0.013	0.01	0.009	0.01	0.01	0.008
Maltese	0.004	0.004	0.005	0.005	0.004	0.006
Nogai	0.005	0.003	0.004	0.004	0.004	0.005
North_Ossetian	0.006	0.005	0.004	0.004	0.005	0.005
Palestinian	0.008	0.007	0.006	0.006	0.005	0.006
Pathan	0.011	0.009	0.008	0.009	0.008	0.008
Romanian	0.002	0.004	0.006	0.006	0.006	0.008
Sardinian	0.01	0.011	0.013	0.013	0.012	0.015
Sicilian	0.003	0.003	0.003	0.004	0.003	0.005
Syrian	0.007	0.005	0.005	0.005	0.005	0.005
Tajik	0.009	0.008	0.008	0.009	0.008	0.008
TSI	0.001	0.002	0.003	0.004	0.003	0.005
Turkmen	0.009	0.007	0.008	0.009	0.008	0.009
Ukrainian	0.003	0.006	0.008	0.009	0.009	0.01
Yemeni	0.01	0.008	0.007	0.008	0.006	0.007

Table S6. Functional annotation and allele frequency distribution of TR variants

			AF in other public databases	
		Novel	Rare (AF < 0.01)	Common (AF ≥ 0.01)
All variants		9,999,451	22,932,246	13,807,782
Functional consequence				
High-confidence pLoFs	Frameshift variant	4,271	3,453	490
	Splice site variant	2,932	3,084	225
	Start loss variant	1		
	Stop gain variant	2,829	5,053	223
	Stop loss variant	4	1	2
Low-confidence pLoFs	Frameshift variant	2,110	2,518	667
	Splice site variant	1,795	3,035	1,784
	Start loss variant	445	949	158
	Stop gain variant	1,221	2,860	345
	Stop loss variant	322	503	149
Missense variants	Deleterious missense	27,086	64,177	2,819
	Other missense	102,360	284,245	41,830
Non-frameshift indels		2,621	6,712	1,728
Synonymous variants		53,768	192,554	41,172
Other effects	Protein-protein contact	149	348	40
	Exon loss variant		2	
	Gene fusion	12	26	7
	Structural interaction variant	3,585	11,044	1,289
	Bidirectional gene fusion	15	36	15
	Transcription Factor Binding Site (TFBS) ablation	116	243	97
	Non-essential splice site variant	22,578	54,573	20,594
	Initiator codon variant	34	60	9
	Stop retained variant	81	180	53
Non-coding variants	Intergenic region	3,617,719	8,210,394	5,318,191
	Intragenic variant	846	1,767	994
	Intron variant	3,538,574	8,102,039	4,871,476
	Upstream gene variant	1,377,873	3,124,448	1,837,118
	TFBS variant	10,101	22,928	10,170
	Sequence feature	70,500	163,949	93,076
	Downstream gene variant	980,383	2,261,777	1,360,894
	Non-coding transcript exon variant	26,319	63,387	38,964
	Untranslated region (UTR) variant	148,801	345,901	163,203

Table S7. Concordance results of the technical replicates

Technical Replicates	Variant Type	Variant Sensitivity	Variant PPV	Variant Specificity	Gentype Concordance	Non-REF Genotype Concordance
Exome/Exome (*n* = 11)	SNV	0.98±0.02	0.97±0.02	0.95±0.03	0.97±0.02	0.97±0.02
	INDEL	0.79±0.18	0.85±0.09	0.84±0.09	0.87±0.16	0.87±0.16
Exome/Genome (*n* = 21)	SNV	0.93±0.09	0.99±0.002	0.98±0.003	0.97±0.007	0.97±0.007
	INDEL	0.61±0.16	0.86±0.04	0.87±0.06	0.85±0.02	0.85±0.02
Genome/Genome (*n* = 6)	SNV	0.99±0.003	0.99±0.0001	0.99±0.0001	0.96±0.005	0.97±0.004
	INDEL	0.89±0.08	0.97±0.003	0.96±0.005	0.76±0.07	0.80±0.07

Dataset S1 (separate file) Y- and mtDNA haplogroups of the TR samples

Dataset S2 (separate file). List of rare homozygous HC-pLoFs

Dataset S3 (separate file). List of common homozygous HC-pLoFs

Dataset S4 (separate file). TR variants that are listed as DM in HGMD

Dataset S5 (separate file). TR variants that are listed as Pathogenic or Pathogenic/Likely pathogenic in ClinVar

Dataset S6 (separate file). Variants in the genes that are causally associated with the phenotypes in the study

SI References

1. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
2. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).
3. C. M. Farrell *et al.*, Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.* **42**, D865-872 (2014).
4. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
5. K. A. Fakhro *et al.*, The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum. Genome Var.* **3**, 16016 (2016).
6. GenomeAsia100KConsortium, The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106-111 (2019).
7. H. Fang *et al.*, Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.* **6**, 89 (2014).
8. R. L. Goldfeder *et al.*, Medical implications of technical accuracy in genome sequencing. *Genome Med.* **8**, 24 (2016).
9. A. Manichaikul *et al.*, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).
10. S. Akbayram *et al.*, The frequency of consanguineous marriage in eastern Turkey. *Genet. Couns.* **20**, 207-214 (2009).
11. A. Belkadi *et al.*, Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6713-6718 (2016).
12. A. Auton *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

13. I. Lazaridis *et al.*, Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419-424 (2016).
14. E. M. Scott *et al.*, Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* **48**, 1071-1076 (2016).
15. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
16. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
17. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer Science & Business Media 2009).
18. Z. Zhang, Reshaping and aggregating data: an introduction to reshape package. *Ann. Transl. Med.* **4**, 78 (2016).
19. H. Wickham, R. François, L. Henry, K. Müller (2018) dplyr: A Grammar of Data Manipulation.
20. H. Wickham (2019) stringr: Simple, Consistent Wrappers for Common String Operations.
21. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome. Res.* **19**, 1655-1664 (2009).
22. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
23. H. Hunter-Zinck *et al.*, Population genetic structure of the people of Qatar. *Am. J. Hum. Genet.* **87**, 17-25 (2010).
24. F. C. Ceballos, S. Hazelhurst, M. Ramsay, Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. *BMC Genomics* **19**, 106 (2018).
25. T. J. Pemberton *et al.*, Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275-292 (2012).
26. R. McQuillan *et al.*, Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359-372 (2008).
27. H. Chen, Y. Lu, D. Lu, S. Xu, Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. *BMC Bioinformatics* **22**, 114 (2021).
28. J. Z. Li *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 (2008).
29. H. Weissensteiner *et al.*, HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58-63 (2016).
30. M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386-394 (2009).
31. C. Cinnioğlu *et al.*, Excavating Y-chromosome haplotype strata in Anatolia. *Hum. Genet.* **114**, 127-148 (2004).
32. D. Comas *et al.*, Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur. J. Hum. Genet.* **12**, 495-504 (2004).
33. P. Lindenbaum, JVarkit: java-based utilities for Bioinformatics. http://dx.doi.org/http://dx.doi.org/10.6084/m9.figshare.1425030.
34. S. E. Hunt *et al.*, Ensembl variation resources. *Database (Oxford)* **2018** (2018).
35. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
36. D. G. MacArthur *et al.*, A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012).
37. B. B. Cummings *et al.*, Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452-458 (2020).
38. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886-D894 (2019).
39. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248-249 (2010).

40. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1-9 (2016).
41. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
42. X. Liu, C. Wu, C. Li, E. Boerwinkle, dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* **37**, 235-241 (2016).
43. P. Sulem *et al.*, Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448-452 (2015).
44. D. Saleheen *et al.*, Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235-239 (2017).
45. V. M. Narasimhan *et al.*, Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474-477 (2016).
46. A. Rausell *et al.*, Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 13626-13636 (2020).
47. P. D. Stenson *et al.*, The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197-1207 (2020).
48. M. J. Landrum *et al.*, ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062-D1067 (2018).
49. B. L. Browning, Y. Zhou, S. R. Browning, A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338-348 (2018).
50. O. Delaneau, J. F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5-6 (2013).
51. B. N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, bay119 (2009).
52. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).

Web Resources

1000 Genomes Project, https://www.internationalgenome.org/
BCFtools, http://samtools.github.io/bcftools/bcftools.html
Haplogrep 2, https://haplogrep.i-med.ac.at/app/
Illumina gvcfgenotyper, https://github.com/Illumina/gvcfgenotyper
IMPUTE2 reference data, https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
International Society of Genetic Genealogy (ISOOG), https://isogg.org/
Isaac Variant Caller, https://github.com/sequencing/isaac_variant_caller
gnomAD, https://gnomad.broadinstitute.org/
mosdepth, https://github.com/brentp/mosdepth)
NHBLI GO Exome Sequencing Project (ESP) Exome variant server, https://evs.gs.washington.edu/EVS/
Samtools, https://github.com/samtools/samtools
The map of Turkey, https://www.yourfreetemplates.com/